(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－215993
（Ｐ２００１－２１５９９３Ａ）
(43)公開日　平成13年８月10日(2001.8.10)

(51)Int.Cl.[7]	識別記号	Ｆ Ｉ			テーマコード（参考）
G10L 15/22		G06F 3/16	320	H	5D015
G06F 3/16	320	G10L 3/00	571	U	5D045
G10L 13/00				R	9A001
15/10			531	N	

審査請求　未請求　請求項の数16　ＯＬ　（全13頁）

(21)出願番号　特願2000－22225(Ｐ2000－22225)

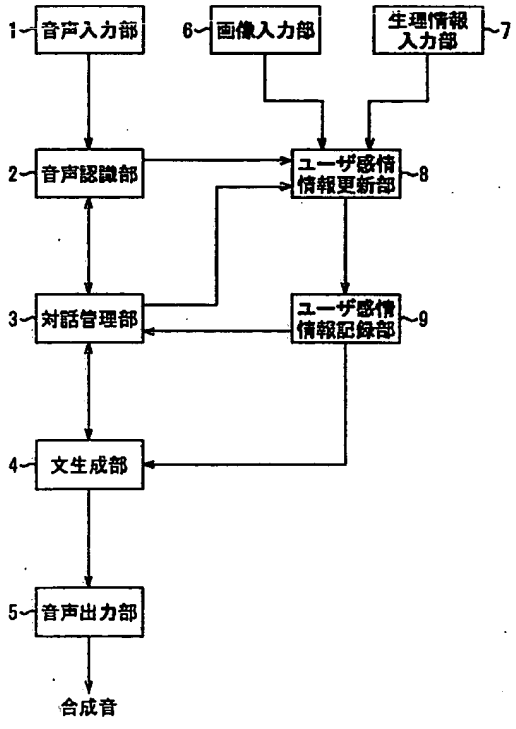
(22)出願日　平成12年１月31日(2000.1.31)

(71)出願人　000002185
ソニー株式会社
東京都品川区北品川６丁目７番35号
(72)発明者　浅野　康治
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　青柳　誠一
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(74)代理人　100082131
弁理士　稲本　義雄

最終頁に続く

(54)【発明の名称】対話処理装置および対話処理方法、並びに記録媒体

(57)【要約】
【課題】　ユーザの感情の状態によって、バリエーションに富んだ対話を行う。
【解決手段】　音声認識部２では、ユーザからの音声が認識されるとともに、その音声の韻律情報が抽出される。対話管理部３では、音声認識部２による音声認識結果に含まれる語句の概念情報が抽出される。画像入力部６では、ユーザの顔が撮像され、顔画像情報が出力される。生理情報入力部７では、ユーザの脈拍数等の生理情報が感知される。そして、ユーザ感情情報更新部８は、上述の韻律情報や、概念情報、顔画像情報、生理情報に基づいて、ユーザの感情を推定し、対話管理部３および文生成部４では、その感情の推定結果に基づいて、ユーザに出力する出力文が生成される。



対話システム

1

【特許請求の範囲】

【請求項１】　ユーザとの対話を行う対話処理装置であって、

ユーザから入力された語句の概念を抽出する概念抽出手段と、

前記ユーザから入力された語句の概念に基づいて、前記ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定手段と、

前記感情情報に基づいて、前記ユーザに出力する出力文を生成する出力文生成手段とを備えることを特徴とする対話処理装置。

【請求項２】　前記感情推定手段は、前記出力文にも基づいて、前記ユーザの感情を推定することを特徴とする請求項１に記載の対話処理装置。

【請求項３】　前記感情推定手段は、前記ユーザを撮像して得られる画像にも基づいて、前記ユーザの感情を推定することを特徴とする請求項１に記載の対話処理装置。

【請求項４】　前記感情推定手段は、前記ユーザの生理現象にも基づいて、前記ユーザの感情を推定することを特徴とする請求項１に記載の対話処理装置。

【請求項５】　外部から得られる音響信号を処理する音響処理手段をさらに備え、

前記感情推定手段は、前記音響処理手段の処理結果にも基づいて、前記ユーザの感情を推定することを特徴とする請求項１に記載の対話処理装置。

【請求項６】　前記ユーザの音声を認識する音声認識手段をさらに備え、

前記概念抽出手段は、前記ユーザの音声の音声認識結果に含まれる語句の概念を抽出することを特徴とする請求項１に記載の対話処理装置。

【請求項７】　前記感情推定手段は、前記ユーザの音声の韻律情報にも基づいて、前記ユーザの感情を推定することを特徴とする請求項６に記載の対話処理装置。

【請求項８】　前記出力文生成手段は、前記感情情報に基づいて、前記出力文の表現を変更することを特徴とする請求項１に記載の対話処理装置。

【請求項９】　前記出力文生成手段は、前記感情情報に基づいて、前記出力文の個数を変更することを特徴とする請求項１に記載の対話処理装置。

【請求項１０】　前記出力文は、相づちを意味するものであることを特徴とする請求項９に記載の対話処理装置。

【請求項１１】　前記感情情報を記憶する記憶手段をさらに備え、

前記出力文生成手段は、前記記憶手段に記憶された前記感情情報に基づいて、前記出力文を生成することを特徴とする請求項１に記載の対話処理装置。

【請求項１２】　前記出力文を出力する出力文出力手段をさらに備えることを特徴とする請求項１に記載の対話処理装置。

2

【請求項１３】　前記出力文出力手段は、前記出力文を合成音で出力することを特徴とする請求項１２に記載の対話処理装置。

【請求項１４】　前記出力文出力手段は、前記感情情報に基づいて、前記合成音の韻律を制御することを特徴とする請求項１３に記載の対話処理装置。

【請求項１５】　ユーザとの対話を行うための対話処理方法であって、

ユーザから入力された語句の概念を抽出する概念抽出ステップと、

前記ユーザから入力された語句の概念に基づいて、前記ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定ステップと、

前記感情情報に基づいて、前記ユーザに出力する出力文を生成する出力文生成ステップとを備えることを特徴とする対話処理方法。

【請求項１６】　ユーザとの対話を行うための対話処理を、コンピュータに行わせるプログラムが記録されている記録媒体であって、

ユーザから入力された語句の概念を抽出する概念抽出ステップと、

前記ユーザから入力された語句の概念に基づいて、前記ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定ステップと、

前記感情情報に基づいて、前記ユーザに出力する出力文を生成する出力文生成ステップとを備えるプログラムが記録されていることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、対話処理装置および対話処理方法、並びに記録媒体に関し、特に、例えば、ユーザの感情を考慮した対話を行うことができるようにする対話処理装置および対話処理方法、並びに記録媒体に関する。

【０００２】

【従来の技術】いわゆる対話システムにおいては、ユーザから入力があると、その入力の意味内容に対応した応答文が生成されて出力される。

【０００３】

【発明が解決しようとする課題】従って、従来の対話システムでは、ユーザの感情がどのような状態であっても、入力の意味内容が同一であれば、同じような応答文が出力され、その結果、同じような対話が行われることになる。

【０００４】本発明は、このような状況に鑑みてなされたものであり、ユーザの感情の状態によって、バリエーションに富んだ対話を行うことができるようにするものである。

【０００５】

【課題を解決するための手段】本発明の対話処理装置は、ユーザから入力された語句の概念を抽出する概念抽出手段と、ユーザから入力された語句の概念に基づいて、ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定手段と、感情情報に基づいて、ユーザに出力する出力文を生成する出力文生成手段とを備えることを特徴とする。

【０００６】感情推定手段には、出力文にも基づいて、ユーザの感情を推定させることができる。

【０００７】また、感情推定手段には、ユーザを撮像して得られる画像にも基づいて、ユーザの感情を推定させることができる。

【０００８】さらに、感情推定手段には、ユーザの生理現象にも基づいて、ユーザの感情を推定させることができる。

【０００９】本発明の対話処理装置には、外部から得られる音響信号を処理する音響処理手段をさらに設けることができ、この場合、感情推定手段には、音響処理手段の処理結果にも基づいて、ユーザの感情を推定させることができる。

【００１０】本発明の対話処理装置には、ユーザの音声を認識する音声認識手段をさらに設けることができ、この場合、概念抽出手段には、ユーザの音声の音声認識結果に含まれる語句の概念を抽出させることができる。

【００１１】感情推定手段には、ユーザの音声の韻律情報にも基づいて、ユーザの感情を推定させることができる。

【００１２】出力文生成手段には、感情情報に基づいて、出力文の表現を変更させることができる。

【００１３】出力文生成手段には、感情情報に基づいて、出力文の個数を変更させることができる。

【００１４】出力文は、相づちを意味するものとすることができる。

【００１５】本発明の対話処理装置には、感情情報を記憶する記憶手段をさらに設けることができ、この場合、出力文生成手段には、記憶手段に記憶された感情情報に基づいて、出力文を生成させることができる。

【００１６】本発明の対話処理装置には、出力文を出力する出力文出力手段をさらに設けることができる。

【００１７】出力文出力手段には、出力文を合成音で出力させることができる。

【００１８】また、出力文出力手段には、感情情報に基づいて、合成音の韻律を制御させることができる。

【００１９】本発明の対話処理方法は、ユーザから入力された語句の概念を抽出する概念抽出ステップと、ユーザから入力された語句の概念に基づいて、ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定ステップと、感情情報に基づいて、ユーザに出力する出力文を生成する出力文生成ステップとを備えることを特徴とする。

【００２０】本発明の記録媒体は、ユーザから入力された語句の概念を抽出する概念抽出ステップと、ユーザから入力された語句の概念に基づいて、ユーザの感情を推定し、その感情を表す感情情報を出力する感情推定ステップと、感情情報に基づいて、ユーザに出力する出力文を生成する出力文生成ステップとを備えるプログラムが記録されていることを特徴とする。

【００２１】本発明の対話処理装置および対話処理方法、並びに記録媒体においては、ユーザから入力された語句の概念が抽出され、その概念に基づいて、ユーザの感情が推定される。そして、その結果得られる感情情報に基づいて、ユーザに出力する出力文が生成される。

【００２２】

【発明の実施の形態】図１は、本発明を適用した対話システム（システムとは、複数の装置が論理的に集合したものをいい、各構成の装置が同一筐体中にあるか否かは問わない）の一実施の形態の構成例を示している。

【００２３】音声入力部１は、例えば、マイク（マイクロフォン）およびアンプ等で構成され、ユーザの音声を、電気信号としての音声信号に変換し、必要に応じて増幅して、その音声信号を、音声認識部２に供給する。

【００２４】音声認識部２は、音声入力部１からの音声信号を音響処理し、さらに、その音響処理結果に基づいて、ユーザの音声を認識する。この音声認識結果は、対話管理部３に供給される。また、音声認識部２は、音声信号を音響処理することにより得られるユーザの音声の韻律情報を、ユーザ感情情報更新部８に供給する。

【００２５】対話管理部３は、ユーザ感情情報記録部９が保持（記憶）している、ユーザの感情を表す感情情報を考慮して、音声認識部２からの音声認識結果に対する応答等としての、ユーザに出力する出力文の内容を生成し、その内容を表す内容情報を、文生成部４に供給する。また、対話管理部３は、音声認識部２からの音声認識結果に含まれる語句や、自身が生成した内容情報に対応する出力文に含まれる語句の概念を抽出し、その概念を表す概念情報を、ユーザ感情情報更新部８に供給する。

【００２６】文生成部４は、ユーザ感情情報記録部９が保持している感情情報を考慮しながら、対話管理部３からの内容情報に対応する、例えばテキストの出力文を生成し、さらに、その出力文に対応する合成音の音声信号を生成して、音声出力部５に供給する。

【００２７】音声出力部５は、例えば、アンプおよびスピーカ等で構成され、文生成部４からの音声信号を、必要に応じて増幅し、スピーカから出力する。

【００２８】画像入力部６は、例えば、レンズ、ＣＣＤ(Charge Coupled Device)、Ａ／Ｄ変換器等で構成され、ユーザの顔等を撮像して、その結果得られる顔画像のディジタルデータ（画像データ）である顔画像情報を、ユーザ感情情報更新部８に供給する。

5

【００２９】生理情報入力部７は、例えば、脈拍計、発汗量や熱を測定するセンサ等で構成され、ユーザの脈拍や、発汗量、熱等の生理的な情報を感知し、その結果得られる生理情報を、ユーザ感情情報更新部８に供給する。

【００３０】ユーザ感情情報更新部８は、音声認識部２からのユーザの音声の韻律情報や、対話管理部３からの音声認識結果等に含まれる語句の概念情報、画像入力部６からの顔画像情報、生理情報入力部７からの生理情報に基づいて、ユーザの感情の状態を推定する。さらに、ユーザ感情情報更新部８は、その推定の結果得られる感情情報によって、ユーザ感情情報記録部９に保持されている感情情報を更新する。

【００３１】ユーザ感情情報記録部９は、ユーザの感情としての、例えば、喜びや、怒り、驚き、悲しみ等の状態を、所定の範囲の数値で表す感情情報を保持している。

【００３２】次に、図２のフローチャートを参照して、図１の対話システムの基本的な処理の流れについて説明する。

【００３３】ユーザにより発話が行われると、音声入力部１は、ステップＳ１において、その発話された音声に対して音声入力処理を施し、その結果得られる音声信号を、音声認識部２に出力する。即ち、音声入力部１は、ユーザの音声を、電気信号としての音声信号に変換し、その音声信号を、必要に応じて増幅して、音声認識部２に供給する。

【００３４】音声認識部２は、ステップＳ２において、音声入力部２からの音声信号に基づいて、ユーザの音声を認識し、その音声認識結果を、対話管理部３に供給する。さらに、音声認識部２は、音声入力部２からの音声信号から、ユーザの音声の韻律情報を抽出し、ユーザ感情情報更新部８に供給する。

【００３５】その後、ステップＳ３に進み、ユーザ感情情報記録部９に保持されている感情情報を更新する準備を行う処理が行われる。

【００３６】即ち、ステップＳ３では、対話管理部３は、音声認識部２からのユーザの音声の音声認識結果等に基づいて、感情情報を更新するのに用いる、上述の概念情報を得る感情情報更新用対話管理処理を行い、その概念情報を、ユーザ感情情報更新部８に供給する。さらに、ステップＳ３では、画像入力部６は、ユーザの顔を撮像して、顔画像情報を得る画像入力処理を行い、その顔画像情報を、ユーザ感情情報更新部８に供給する。また、ステップＳ３では、生理情報入力部７は、ユーザの生理情報を得る生理情報入力処理を行い、その生理情報を、ユーザ感情情報更新部８に供給する。

【００３７】ユーザ感情情報更新部８は、ステップＳ４において、音声認識部２からのユーザの音声の韻律情報や、対話管理部３からの概念情報、画像入力部６からの顔画像情報、生理情報入力部７からの生理情報に基づいて、ユーザの感情の状態を推定する。さらに、ステップＳ４では、ユーザ感情情報更新部８は、その推定の結果得られる感情情報によって、ユーザ感情情報記録部９に保持されている感情情報を更新する。

【００３８】その後、ステップＳ５において、対話管理部３は、ユーザ感情情報記録部９が保持（記憶）している、ユーザの感情を表す感情情報を考慮して、音声認識部２からの音声認識結果に対する応答等としての、ユーザに出力する出力文の内容を表す内容情報を生成する文生成用対話管理処理を行い、その内容情報を、文生成部４に供給する。

【００３９】そして、ステップＳ６において、文生成部４は、ユーザ感情情報記録部９が保持している感情情報を考慮しながら、対話管理部３からの内容情報に対応するテキストの出力文を生成し（文生成処理を行い）、さらに、その出力文に対応する合成音の音声信号を生成して、音声出力部５に供給する。

【００４０】音声出力部５は、ステップＳ７において、文生成部４からの音声信号を増幅し、スピーカから出力する音声出力処理を行い、処理を終了する。

【００４１】なお、上述の場合には、対話システムにおいて、ユーザが何らかの発話を行ったことをトリガとして、合成音の出力（以下、適宜、対話システムの発話ともいう）が行われるから、その合成音は、ユーザの発話に対する応答となるが、対話システムにおいては、ユーザの発話以外をトリガとして、発話を行うようにすることも可能である。

【００４２】即ち、対話システムにおいては、例えば、所定の時間ごとに発話を行うようにすることが可能である。また、例えば、画像入力部６において、ユーザの顔画像が得られたとき（単に、顔画像が得られたときの他、所定の表情の顔画像が得られたときも含む）や、生理情報入力部７において、所定の生理情報が得られたときに、発話を行うようにすることも可能である。さらに、例えば、ユーザ感情情報記録部９に保持された感情情報が所定の値以上または以下になったときに、発話を行うようにすることも可能である。これらの場合は、対話システムが、ユーザに話しかけ、その応答をユーザが返す形で、対話が行われることになる。

【００４３】次に、図３は、図１の音声認識部２の構成例を示している。

【００４４】音声入力部１からの音声信号は、ＡＤ(Analog Digtal)変換部１１に供給されるようになっており、ＡＤ変換部１１は、その音声信号を、アナログ信号からディジタル信号に変換し、その結果得られる音声データを、特徴抽出部１２に供給する。特徴抽出部１２は、ＡＤ変換部１１からの音声データについて、適当なフレームごとに音響処理を施すことで、例えば、スペクトルや、線形予測係数、ケプストラム係数、線スペクト

ル対、ＭＦＣＣ(Mel Frequency Cepstrum Coefficient)等の特徴パラメータを抽出し、マッチング部１３に供給する。

【００４５】また、特徴抽出部１２は、音声データに音響処理を施すことにより得られる、例えば、発話速度や、ピッチ周波数、パワー等の韻律情報を、ユーザ感情情報更新部８に供給する。なお、発話速度としては、例えば、１フレームあたりのモーラ数等を用いることができる。

【００４６】マッチング部１３は、特徴抽出部１２から供給される特徴パラメータに基づき、音響モデルデータベース１４、辞書データベース１５、および文法データベース１６を必要に応じて参照しながら、ユーザの音声（入力音声）を認識する。

【００４７】即ち、音響モデルデータベース１４は、音声認識する音声の言語における個々の音素や音節などの音響的な特徴を表す音響モデルを記憶している。ここで、音響モデルとしては、例えば、HMM(Hidden Markov Model)などを用いることができる。辞書データベース１５は、認識対象の各単語について、その発音に関する情報が記述された単語辞書を記憶している。文法データベース１６は、辞書データベース１５の単語辞書に登録されている各単語が、どのように連鎖する（つながる）かを記述した文法規則を記憶している。ここで、文法規則としては、例えば、文脈自由文法（ＣＦＧ）やＨＰＳＧ（Head-driven Phrase Structure Grammar）（主辞駆動句構造文法）、統計的な単語連鎖確率（Ｎ－ｇｒａｍ）などに基づく規則を用いることができる。

【００４８】マッチング部１３は、辞書データベース１５の単語辞書を参照することにより、音響モデルデータベース１４に記憶されている音響モデルを接続することで、単語の音響モデル（単語モデル）を構成する。さらに、マッチング部１３は、幾つかの単語モデルを、文法データベース１６に記憶された文法規則を参照することにより接続し、そのようにして接続された単語モデルを用いて、特徴パラメータに基づき、例えば、HMM法等によって、ユーザの音声を認識する。

【００４９】そして、マッチング部１３による音声認識結果としての音韻情報は、例えば、テキストやワードグラフ等で、対話管理部３に出力される。

【００５０】次に、図４は、図１の対話管理部３の構成例を示している。

【００５１】音声認識部２が出力するユーザの音声認識結果は、言語処理部２１に供給されるようになっており、言語処理部２１は、シソーラスデータベース２３や、言語処理用データベース２４、履歴データベース２５を必要に応じて参照しながら、音声認識結果を処理し、その音声認識結果が表す意味や概念を、対話処理部２２に供給する。

【００５２】即ち、シソーラスデータベース２３には、単語を、その概念によって階層構造に分類したシソーラスが記憶されており、言語処理部２１は、このシソーラスを参照することにより、音声認識結果に含まれる単語の概念を認識する。

【００５３】ここで、シソーラスとしては、例えば、国立国語研究所によって発表されている分類語彙表等を用いることができる。

【００５４】言語処理用データベース２４には、各単語の表記や必要な品詞情報などが記述された単語辞書と、その単語辞書に記述された各単語の情報に基づいて、単語連鎖に関する制約等が記述された構文／意味規則が記憶されており、言語処理部２１は、その単語辞書や構文／意味規則に基づいて、そこに入力される音声認識結果の形態素解析を行う。さらに、言語処理部２１は、その形態素解析結果に基づいて、音声認識結果の構文解析、さらには、その意味内容の理解を行う。そして、言語処理部２１は、以上のようにして得られる音声認識結果を構成する各単語の概念や、音声認識結果の意味内容の理解の結果（以下、適宜、まとめて言語処理結果という）を、対話処理部２２に出力する。

【００５５】ここで、言語処理部２１では、例えば、正規文法や、文脈自由文法、ＨＰＳＧ、統計的な単語連鎖確率を用いて、構文解析や意味内容の理解を行うことができる。

【００５６】また、言語処理部２１は、必要に応じて、履歴データベース２５も参照しながら処理を行う。即ち、履歴データベース２５には、ユーザが発話した音声の音声認識結果と、その発話に対して、対話システムが出力した応答との組や、あるいは、対話システムの出力と、その出力に対して、ユーザが発話した音声の音声認識結果との組等の形で、ユーザと対話システムとの間の対話の履歴（対話履歴）が記憶されるようになっており、言語処理部２１は、対話履歴を参照することで、音声認識結果における主語等の省略や、照応表現等の解析を行い、これにより、例えば、ユーザの音声認識結果に含まれる代名詞が、具体的に何を意味しているのか等を認識するようになっている。

【００５７】なお、シソーラスデータベース２３および言語処理用データベース２４に記憶されている情報は、基本的には更新されないから、いわば静的な情報ということができる。これに対して、履歴データベース２５に記憶されている対話履歴は、ユーザにより発話が行われ、あるいは、対話システムが、ユーザに対して何らかの出力を行うと、後述する対話処理部２２によって更新されていくので、いわば動的な情報ということができる。

【００５８】上述したように、言語処理部２１は、シソーラスデータベース２３を参照することで、音声認識結果を構成する各単語（語彙）の概念を抽出するが、その概念が、感情を表すものであるとき、その感情を表す概

念を、概念情報として、ユーザ感情情報更新部８に供給する。即ち、言語処理部２１は、シソーラス上において、例えば、「喜び」や、「怒り」、「驚き」、「悲しみ」、「苦しさ」、「恥ずかしさ」、「楽しさ」等の、感情を表す概念に属する単語が、音声認識結果に含まれるとき、その概念を表す概念情報を、ユーザ感情情報更新部８に供給する。

【００５９】なお、言語処理部２１は、音声認識結果に含まれる単語の概念情報の他、対話履歴として記憶されている対話システムの出力に含まれる単語の概念情報も、必要に応じて抽出し、ユーザ感情情報更新部８に供給するようになっている。

【００６０】即ち、ユーザ感情情報更新部８は、上述のように、ユーザの感情の状態を推定するが、その推定にあたっては、音声認識結果に含まれる単語の概念情報は勿論であるが、対話システムの出力に含まれる単語の概念情報も役に立つ場合がある。具体的には、例えば、対話システムにおいて、ユーザを愚弄するような発話を行った場合には、ユーザが怒ることが予想される。このため、言語処理部２１は、対話履歴として記憶されている対話システムの出力に含まれる単語の概念情報も、シソーラスを参照することで抽出し、音声認識結果に含まれる単語の概念情報とともに、ユーザ感情情報更新部８に供給するようになっている。

【００６１】対話処理部２２は、言語処理部２１からの言語処理結果、およびユーザ感情情報記録部９に保持されている、ユーザの感情の状態を表す感情情報に基づき、履歴データベース２５やシナリオデータベース２６を参照しながら、ユーザの音声認識結果に対する応答等としての、ユーザに出力する出力文の内容を生成し、その内容を表す内容情報を、文生成部４に供給する。

【００６２】即ち、シナリオデータベース２６は、例えば、ユーザとの対話パターンの規則としてのシナリオを、タスク（話題）ごとに記憶しており、対話処理部２２は、基本的には、シナリオデータベース２６に記憶されているシナリオの中から、言語処理部２１からの言語処理結果に基づいて、ユーザとの対話に用いるものを決定し、そのシナリオにしたがって、内容情報を生成する。

【００６３】具体的には、例えば、ビデオ予約等の目的志向型のタスクについては、例えば、次のようなシナリオが記憶されている。

【００６４】

```
(action(Question(date,start_time,end_time,channel)))
        (date   ???)            #日付
        (start_time          ???)#開始時刻
        (end_time            ???)  #終了時刻
        (channel             ???)    #チャンネル
```

・・・（１）

【００６５】ここで、（１）のシナリオによれば、言語処理部２１による言語処理結果が、録画の要求を表すものである場合には、対話処理部２２において、録画を行う日付、録画を開始する時刻、録画を終了する時刻、録画を行うチャンネルを、そのような順番で質問する旨の内容情報が生成される。

【００６６】また、例えば、無目的型の対話（いわゆる雑談）を行うためのシナリオとしては、次のようなものが記憶されている。

【００６７】

```
If X exist then speak (Y)    # X:キーワード, Y:応答文
(お金　何が欲しいの)    # (X Y)
(食べたい　お腹がすいているの)
```

・・・（２）

【００６８】ここで、（２）のシナリオによれば、言語処理部２１による言語処理結果に、「お金」というキーワードが含まれていれば、対話処理部２２において、「何が欲しいの」という、質問を行う旨の内容情報が生成される。また、言語処理部２１による言語処理結果に、「食べたい」というキーワードが含まれていれば、対話処理部２２において、「お腹がすいているの」という、質問を行う旨の内容情報が生成される。

【００６９】また、対話処理部２２は、例えば、言語処理部２１からの言語処理結果だけでなく、ユーザ感情情報記録部９に保持されている感情情報にも基づいて、使用するシナリオを決定する。即ち、例えば、言語処理部２１からの言語処理結果が、ユーザが挨拶をしたことを表している場合において、感情情報が、「楽しさ」や「うれしさ」が通常レベルであることを表しているときには、あるいは、「怒り」や「いらつき」が大であることを表しているときには、対話処理部２２は、ユーザに「こんにちは」と、単に挨拶を返すシナリオの使用を決定する。また、例えば、例えば、言語処理部２１からの言語処理結果が、ユーザが挨拶をしたことを表している場合において、感情情報が、「楽しさ」や「うれしさ」が大であることを表しているときには、対話処理部２２は、ユーザに「何か良いことがあったのですか？」と問い合わせるシナリオの使用を決定する。

【００７０】なお、シナリオデータベース２６には、シナリオの他、ユーザと対話を行うにあたっての一般的な

知識も記憶されている。即ち、シナリオデータベース２６には、例えば、言語処理部２１による言語処理結果が、ユーザが挨拶をしたことを表している場合には、その挨拶に対する挨拶を行うことを指示する情報が、一般的な知識として記憶されている。また、シナリオデータベース２６には、例えば、雑談時に使用する話題（トピックス）なども、一般的な知識として記憶されている。

【００７１】さらに、対話処理部２２は、言語処理部２１からの言語処理結果や、自身が生成した内容情報、さらには、その内容情報を生成するのに用いたシナリオに関する情報等を、対話履歴として、履歴データベース２５に記憶させる。

【００７２】また、対話処理部２２は、必要に応じて、対話履歴を参照し、これにより、例えば、音声認識結果や、その意味の理解に誤りがあったことが、後から判明した場合等に対処するようにもなっている。

【００７３】次に、図５は、図１の文生成部４の構成例を示している。

【００７４】テキスト文生成部３１には、対話管理部３から内容情報が供給されるようになっており、テキスト文生成部３１は、必要に応じて、辞書データベース３４および生成文法データベース３５を参照しながら、内容情報に対応する（即した）、テキストの出力文を生成する。

【００７５】即ち、辞書データベース３４には、各単語の品詞情報や、読み、アクセント等の情報が記述された単語辞書が記憶されており、生成用文法データベース３５には、出力文の例のテンプレート、さらには、出力文を生成するのに必要な単語の活用規則や、語順の制約情報等の生成用文法規則が記憶されている。そして、テキスト文生成部３１は、内容情報に対応するテンプレートを選択し、さらに、必要な単語を単語辞書から選択する。さらに、テキスト文生成部３１は、生成用文法規則を参照して、語尾等を適切に変えながら、単語をテンプレートにあてはめることで、内容情報に対応する出力文を生成する。

【００７６】また、テキスト文生成部３１には、ユーザ感情情報記録部９に保持されている感情情報も供給されるようになっており、テキスト文生成部３１は、その感情情報に基づいて、出力文の表現を変更する。即ち、生成用文法データベース３５には、同一内容のテンプレートで、表現の異なるものが記憶されており、テキスト文生成部３１は、そのような同一内容のテンプレートから、所定の表現のものを、感情情報に基づいて選択する。また、テキスト文生成部３１は、テンプレートにあてはめる単語についても、所定の表現のものを、感情情報に基づいて選択する。さらに、テキスト文生成部３１は、語尾等の変更も、感情情報に基づいて行う。

【００７７】これにより、例えば、感情情報が、「怒り」や「悲しみ」のレベルが大であることを表しているときには、テキスト生成部３１において、比較的丁寧な表現の出力文が生成される。また、例えば、感情情報が、「楽しさ」や「喜び」のレベルが大であることを表しているときには、テキスト生成部３１において、いわゆるラフな表現の出力文が生成される。

【００７８】なお、出力文の生成の方法としては、テンプレートを用いる方法の他、例えば、格構造に基づく方法等を採用することも可能である。

【００７９】テキスト文生成部３１は、出力文を生成すると、その形態素解析や構文解析等を行い、後段の規則合成部３２で行われる規則音声合成に必要な情報を抽出する。ここで、規則音声合成に必要な情報としては、例えば、ポーズの位置や、アクセントおよびイントネーションを制御するための情報その他の韻律情報や、各単語の発音等の音韻情報などがある。

【００８０】テキスト文生成部３１で得られた情報は、規則合成部３２に供給され、規則合成部３２では、音素片データベース３６を用いて、テキスト文生成部３１において生成された出力文に対応する合成音の音声データ（ディジタルデータ）が生成される。

【００８１】即ち、音素片データベース３６には、例えば、ＣＶ(Consonant, Vowel)や、ＶＣＶ、ＣＶＣ等の形で音素片データが記憶されており、規則合成部３２は、テキスト文生成部３１からの情報に基づいて、必要な音素片データを接続し、さらに、ポーズ、アクセント、イントネーション等を適切に付加することで、テキスト文生成部３１で生成された出力文に対応する合成音の音声データを生成する。

【００８２】また、規則合成部３２には、ユーザ感情情報記録部９に保持されている感情情報が供給されるようになっており、規則合成部３２は、感情情報に基づいて、接続された音素片データに付加するポーズや、アクセント、イントネーション、さらには、発話速度、ピッチ周波数等の韻律情報を制御する。即ち、これにより、規則合成部３２では、例えば、感情情報が、ユーザが興奮していることを表しているときには、ゆっくりとした、落ち着いた調子の合成音の音声データが生成される。また、例えば、感情情報が、ユーザが楽しそうであることを表しているときには、規則合成部３２では、やはり、楽しそうな調子の合成音の音声データが生成される。

【００８３】なお、感情と音声との関係については、例えば、前川、「音声によるパラ言語情報の伝達；言語学の立場から」、日本音響学会平成９年度秋季研究発表会講演論文集１－３－１０、pp.381-384、平成９年９月等に、その詳細が記載されている。

【００８４】規則合成部３２で得られた合成音の音声データは、ＤＡ(Digital Analog)変換部３３に供給され、そこで、アナログ信号としての音声信号に変換される。この音声信号は、音声出力部５に供給され、これによ

り、テキスト文生成部３１で生成された出力文に対応する合成音が出力される。

【００８５】次に、図６は、図１のユーザ感情情報更新部８の構成例を示している。

【００８６】音声認識部２が出力する韻律情報は韻律情報処理部４１に、対話管理部３が出力する概念情報は概念情報処理部４２に、画像入力部６が出力する顔画像情報は画像情報処理部４３に、生理情報入力部７が出力する生理情報は生理情報処理部４４に、それぞれ供給されるようになっている。

【００８７】韻律情報処理部４１は、そこに供給される韻律情報を処理することにより、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。

【００８８】なお、ユーザの音声の韻律情報から、そのユーザの感情を推定する方法としては、例えば、特開平１０－５５１９４号公報に記載されているもの等を用いることが可能である。

【００８９】概念情報処理部４２は、そこに供給される概念情報を処理することにより、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。即ち、概念情報処理部４２は、概念情報に基づき、「喜び」や「怒り」等といった各感情を表す概念に属する単語が、ユーザと対話システムとの対話において出現した出現頻度をカウントする。そして、概念情報処理部４２は、その出現頻度に基づいて、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。

【００９０】画像情報処理部４３は、そこに供給される顔画像情報を処理することにより、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。

【００９１】即ち、図７は、図６の画像情報処理部４３の構成例を示している。

【００９２】顔画像情報は、特徴抽出部５１に供給され、特徴抽出部５１は、その顔画像情報の特徴量を抽出する。即ち、特徴抽出部５１は、例えば、顔画像情報をウェーブレット(Wavelet)変換し、空間周波数成分を表す係数をコンポーネントとする特徴ベクトルを得て、ベクトル量子化部５２に供給する。

【００９３】ベクトル量子化部５２は、コードブックデータベース５４に記憶されたコードブックにしたがって、特徴抽出部５１からの特徴ベクトルをベクトル量子化し、これにより、１次元のシンボル（列）を得る。

【００９４】即ち、コードブックデータベース５４には、喜んでいる状態や、怒っている状態、驚いている状態、悲しんでいる状態等の、各感情の状態における顔の画像を用いて学習を行うことにより得られたコードブックが記憶されている。なお、ここでは、量子化精度を高めるために、例えば、喜び用コードブックや怒り用コードブックのように、各感情ごとのコードブックが作成されて記憶されている。

【００９５】そして、ベクトル量子化部５２は、コードブックデータベース５４に記憶された各感情ごとのコードブックにしたがって、特徴抽出部５１からの特徴ベクトルをベクトル量子化し、シンボル（コードブックのコードベクトルに割り当てられたコード）を得て、マッチング部５３に出力する。従って、マッチング部５３には、各感情ごとのベクトル量子化結果としてのシンボルが供給される。

【００９６】マッチング部５３は、ベクトル量子化部５２からのシンボルを用い、HMMデータベース５５を参照して、顔画像情報が、例えば、喜んでいる状態、怒っている状態、驚いている状態、悲しんでいる状態等のうちのいずれの感情の状態における顔のものであるかのマッチングを行う。

【００９７】即ち、HMMデータベース５５には、喜んでいる状態や、怒っている状態、驚いている状態、悲しんでいる状態等の、各感情の状態における顔の画像を用いて学習を行うことにより得られた、各感情における顔についてのモデル（HMM）が記憶されている。

【００９８】そして、マッチング部５３は、ベクトル量子化部５２から得られるシンボル系列が観測される確率が最も高いモデルを、ビタビ法により求める。さらに、マッチング部５３は、そのモデルに対応する感情を、ユーザの感情として推定し、その推定結果としての感情情報を、更新処理部４５に出力する。

【００９９】ここで、マッチング部５３において、ベクトル量子化部５２から得られるシンボル系列が観測される確率の計算は、各感情ごとに行われる。即ち、例えば、喜び用コードブックを用いてベクトル量子化を行うことにより得られたシンボル系列が観測される確率の計算は、喜んでいる状態の顔の画像を用いて学習が行われたHMM（喜び用HMM）を用いて行われる。また、例えば、怒り用コードブックを用いてベクトル量子化を行うことにより得られたシンボル系列が観測される確率の計算は、怒っている状態の顔の画像を用いて学習が行われたHMM（怒り用HMM）を用いて行われる。

【０１００】なお、上述のようにして、顔画像情報から、感情を推定する方法については、例えば、坂口、大谷、岸野、「隠れマルコフモデルによる顔動画像からの表情認識」、テレビジョン学会誌、VOL.49, no.8, pp.1060-1067, 1995年8月等に、その詳細が記載されている。

【０１０１】また、顔画像情報から、感情を推定する方法としては、その他、例えば、坂口、森島、「空間周波数情報に基づく基本表情の実時間認識」、第２回知能情報メディアシンポジウム論文集, pp.75-82, １９９６年１２月等に記載されている方法を採用することも可能である。

【０１０２】図６に戻り、生理情報処理部４４は、そこに供給される生理情報を処理することにより、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。ここで、生理情報から、ユーザの感情を推定する方法としては、例えば、各感情と、脈拍数や発汗量等の生理情報との相関を表す関数を、あらかじめ統計的に求めておき、その関数を用いて行う方法等がある。

【０１０３】更新処理部４５は、韻律情報処理部４１、概念情報処理部４２、画像情報処理部４３、および生理情報処理部４４からの感情情報を総合的に用いて、ユーザ感情情報記録部９に保持されている感情情報を更新する最終的な更新値を求め、その更新値によって、ユーザ感情情報記録部９の感情情報を更新する。即ち、更新処理部４５は、例えば、韻律情報処理部４１、概念情報処理部４２、画像情報処理部４３、生理情報処理部４４それぞれからの、各感情に対応する感情情報を重み付け加算して正規化することで、各感情に対応する最終的な感情情報を算出する。そして、更新処理部４５は、この最終的な感情情報によって、ユーザ感情情報記録部９の感情情報を更新する。

【０１０４】ここで、図８は、ユーザ感情情報記録部９が保持している感情情報を示している。各感情に対応する感情情報は、その感情の度合いを、例えば、０乃至１の範囲の実数で表すもので、値が大きいほど、その感情が強い（値が小さいほど、その感情が弱い）ことを示す。更新処理部４５では、このような感情情報としての値が、各感情ごとに更新される。

【０１０５】次に、図９のフローチャートを参照して、図６のユーザ感情情報更新部８の処理（感情情報更新処理）について説明する。

【０１０６】まず最初に、ステップＳ１１において、韻律情報処理部４１、概念情報処理部４２、画像情報処理部４３、および生理情報処理部４４は、上述したようにして、ユーザの感情を推定し、その推定結果としての感情情報を、更新処理部４５に出力する。

【０１０７】更新処理部４５は、ステップＳ１２において、韻律情報処理部４１、概念情報処理部４２、画像情報処理部４３、および生理情報処理部４４からの感情情報を総合的に用いて、ユーザ感情情報記録部９に保持されている感情情報を更新する最終的な更新値を求め、ステップＳ１３に進み、その更新値によって、ユーザ感情情報記録部９の感情情報を更新して、処理を終了する。

【０１０８】次に、上述した一連の処理は、専用のハードウェアにより行うこともできるし、ソフトウェアにより行うこともできる。一連の処理をソフトウェアによって行う場合には、そのソフトウェアを構成するプログラムが、汎用のコンピュータ等にインストールされる。

【０１０９】そこで、図１０は、上述した一連の処理を実行するプログラムがインストールされるコンピュータの一実施の形態の構成例を示している。

【０１１０】プログラムは、コンピュータに内蔵されている記録媒体としてのハードディスク１０５やＲＯＭ１０３に予め記録しておくことができる。

【０１１１】あるいはまた、プログラムは、フロッピーディスク、CD-ROM(Compact Disc Read Only Memory), MO(Magneto optical)ディスク, DVD(Digital Versatile Disc)、磁気ディスク、半導体メモリなどのリムーバブル記録媒体１１１に、一時的あるいは永続的に格納（記録）しておくことができる。このようなリムーバブル記録媒体１１１は、いわゆるパッケージソフトウエアとして提供することができる。

【０１１２】なお、プログラムは、上述したようなリムーバブル記録媒体１１１からコンピュータにインストールする他、ダウンロードサイトから、ディジタル衛星放送用の人工衛星を介して、コンピュータに無線で転送したり、LAN(Local Area Network)、インターネットといったネットワークを介して、コンピュータに有線で転送し、コンピュータでは、そのようにして転送されてくるプログラムを、通信部１０８で受信し、内蔵するハードディスク１０５にインストールすることができる。

【０１１３】コンピュータは、CPU(Central Processing Unit)１０２を内蔵している。CPU１０２には、バス１０１を介して、入出力インタフェース１１０が接続されており、CPU１０２は、入出力インタフェース１１０を介して、ユーザによって、キーボードやマウス等で構成される入力部１０７が操作されることにより指令が入力されると、それにしたがって、ROM(Read Only Memory)１０３に格納されているプログラムを実行する。あるいは、また、CPU１０２は、ハードディスク１０５に格納されているプログラム、衛星若しくはネットワークから転送され、通信部１０８で受信されてハードディスク１０５にインストールされたプログラム、またはドライブ１０９に装着されたリムーバブル記録媒体１１１から読み出されてハードディスク１０５にインストールされたプログラムを、RAM(Random Access Memory)１０４にロードして実行する。これにより、CPU１０２は、上述したフローチャートにしたがった処理、あるいは上述したブロック図の構成により行われる処理を行う。そして、CPU１０２は、その処理結果を、必要に応じて、例えば、入出力インタフェース１１０を介して、LCD(Liquid CryStal Display)やスピーカ等で構成される出力部１０６から出力、あるいは、通信部１０８から送信、さらには、ハードディスク１０５に記録等させる。

【０１１４】ここで、本明細書において、コンピュータに各種の処理を行わせるためのプログラムを記述する処理ステップは、必ずしもフローチャートとして記載された順序に沿って時系列に処理する必要はなく、並列的あるいは個別に実行される処理（例えば、並列処理あるいはオブジェクトによる処理）も含むものである。

【０１１５】また、プログラムは、１のコンピュータにより処理されるものであっても良いし、複数のコンピュータによって分散処理されるものであっても良い。さらに、プログラムは、遠方のコンピュータに転送されて実行されるものであっても良い。

【０１１６】以上のように、少なくとも、ユーザの音声認識結果に含まれる語句の概念に基づいて、ユーザの感情を推定するようにしたので、比較的精度良く、ユーザの感情を推定することができる。さらに、その他、韻律情報や、顔画像情報、生理情報にも基づいて、ユーザの感情を推定するようにしたので、より精度良く、ユーザの感情を推定することができる。さらに、そのような感情の推定結果に基づいて、出力文を生成するようにしたので、ユーザの感情の状態によって、バリエーションに富んだ出力文を、ユーザに提供することが可能となる。

【０１１７】なお、本実施の形態では、音声入力部１に入力された音（音声）について、音声認識を行うようにしたが、音声入力部１に入力された音については、音声認識を行わずに、例えば、その音が、机を叩いている音であるとか、ユーザの息づかいであるといったことを検出し、その検出結果に基づいて、ユーザの感情を推定することも可能である。即ち、例えば、机を叩いていることが連続して検出された場合には、ユーザが怒っていることを推定することができる。また、例えば、息づかいが荒いことが検出された場合には、ユーザが興奮していることを推定することができる。そして、この場合、そのような推定結果に基づいて、「怒り」や「興奮」を表す感情情報の値を大きくするような、アドホック(ad hoc)な更新ルールを適用することができる。

【０１１８】さらに、対話管理部３においては、感情状態に応じて、出力文の生成回数を制御することにより、ユーザに対する発話の回数を変化させることが可能である。具合的には、例えば、ユーザが楽しそうな状態にある場合には、例えば、相づちの回数を増やしたり、その他、対話システムからの発話回数を増やして、積極的に、ユーザとの対話を行うようにすることが可能である。また、例えば、ユーザが悲しそうな状態にある場合には、対話システムからの発話回数を減らして、ユーザに煩わしさを感じさせないようにすることが可能である。

【０１１９】また、本実施の形態では、ユーザからの音声を音声認識し、その音声認識結果に対する応答としての発話を行うようにしたが、その他、例えば、ユーザがキーボードを操作することにより入力される文に対して、応答を行うようにすることも可能である。

【０１２０】さらに、本実施の形態では、ユーザに対する応答等を、合成音で出力するようにしたが、その他、例えば、テキスト等で表示するようにすることも可能である。

【０１２１】また、本発明は、例えば、ディスプレイに表示される仮想的なキャラクタや、あるいは実体のあるロボット等とユーザとの間のユーザインタフェースとして用いることが可能である。この場合、ユーザに対する応答等として、上述したように合成音を出力する他、仮想的なキャラクタの表示状態を変えたり、ロボットに所定の動作を行わせることで、マルチモーダルなインタフェースを実現することができる。

【０１２２】

【発明の効果】本発明の対話処理装置および対話処理方法、並びに記録媒体によれば、ユーザから入力された語句の概念が抽出され、その概念に基づいて、ユーザの感情が推定される。そして、その結果得られる感情情報に基づいて、ユーザに出力する出力文が生成される。従って、ユーザの感情の状態によって、例えば、バリエーションに富んだ対話を行うことが可能となる。

【図面の簡単な説明】

【図１】本発明を適用した対話システムの一実施の形態の構成例を示すブロック図である。

【図２】図１の対話システムの処理を説明するためのフローチャートである。

【図３】図１の音声認識部２の構成例を示すブロック図である。

【図４】図１の対話管理部３の構成例を示すブロック図である。

【図５】図１の文生成部４の構成例を示すブロック図である。

【図６】図１のユーザ感情情報更新部８の構成例を示すブロック図である。

【図７】図６の画像情報処理部４３の構成例を示すブロック図である。

【図８】感情情報を示す図である。

【図９】図６のユーザ感情情報更新部８の処理を説明するためのフローチャートである。

【図１０】本発明を適用したコンピュータの一実施の形態の構成例を示すブロック図である。

【符号の説明】

１　音声入力部，　２　音声認識部，　３　対話管理部，　４　文生成部，５　音声出力部，　６　画像入力部，　７　生理情報入力部，　８　ユーザ感情情報更新部，　９　ユーザ感情情報記録部，　１１　ＡＤ変換部，　１２　特徴抽出部，　１３　マッチング部，　１４　音響モデルデータベース，　１５　辞書データベース，　１６　文法データベース，　２１　言語処理部，　２２　対話処理部，　２３　シソーラスデータベース，　２４　言語処理用データベース，　２５　履歴データベース，　２６　シナリオデータベース，　３１　テキスト文生成部，　３２　規則合成部，　３３　ＤＡ変換部，　３４　辞書データベース，　３５　生成用文法データベース，　３６　音素片データベース，　４１　韻律情報処理部，　４２　概念情報処理部，　４３　画

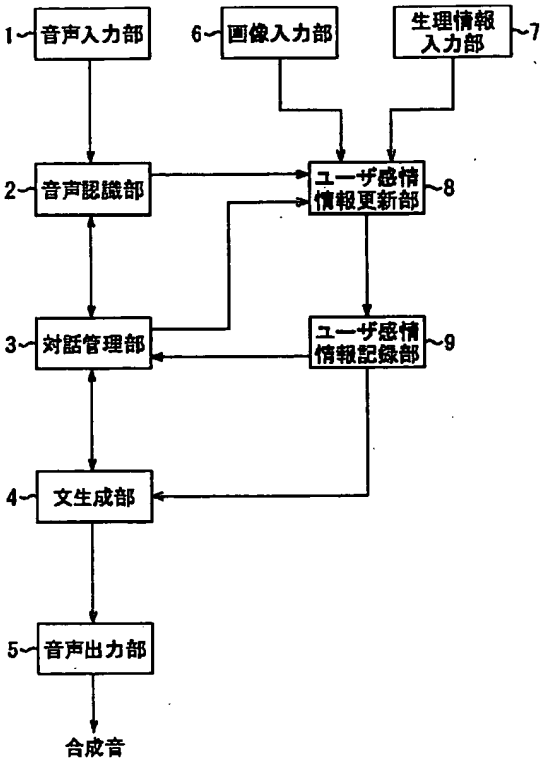
19

像情報処理部，　４４生理情報処理部，　５１　特徴抽出部，　５２　ベクトル量子化部，　５３マッチング部，　５４　コードブックデータベース，　５５　HMMデータベース，　１０１　バス，　１０２　CPU，

20

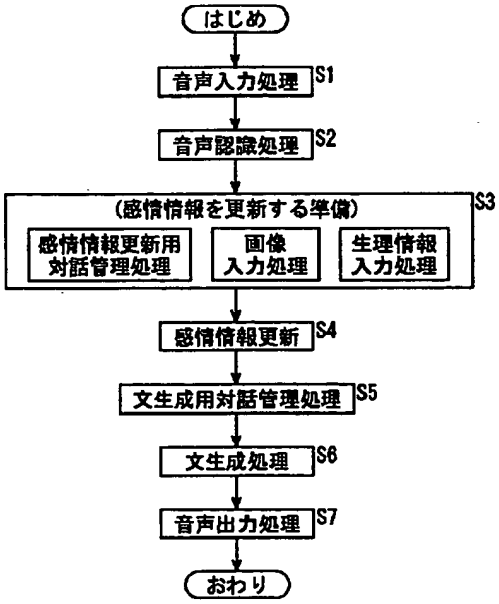
１０３　ROM，　１０４　RAM，　１０５　ハードディスク，　１０６　出力部，　１０７　入力部，　１０８　通信部，　１０９　ドライブ，　１１０　入出力インタフェース，　１１１　リムーバブル記録媒体

【図１】

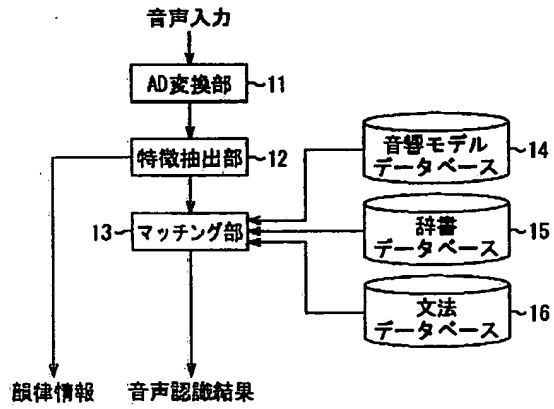


対話システム

【図２】

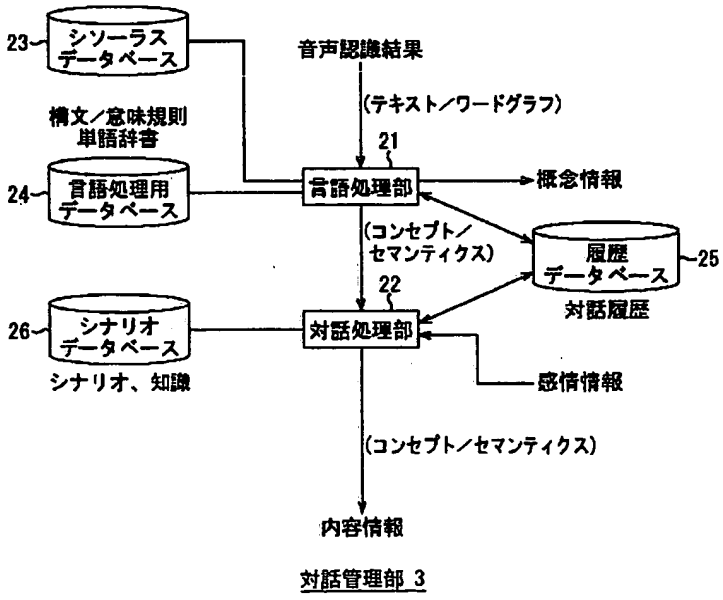


【図３】



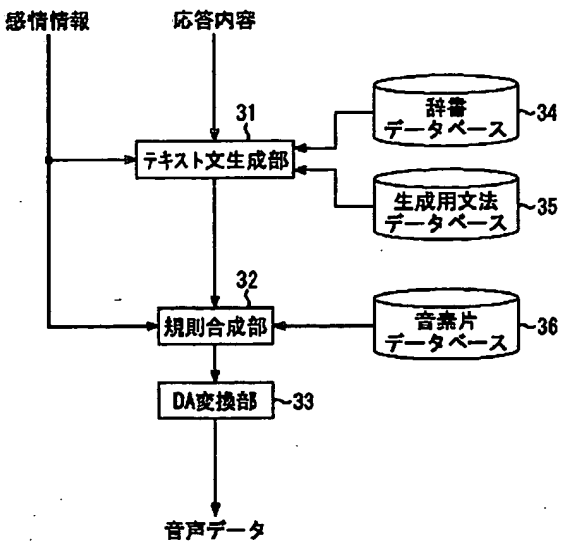
音声認識部 2

【図４】



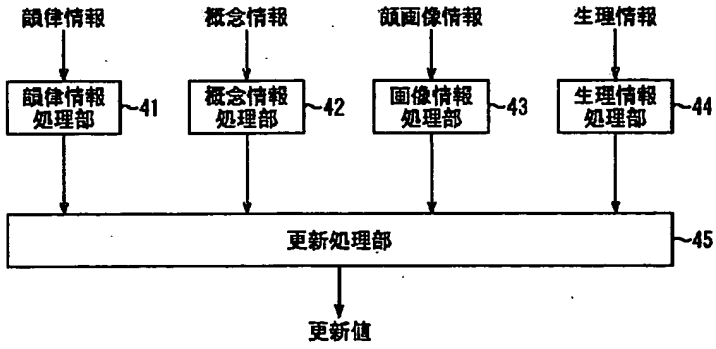
対話管理部 3

【図５】



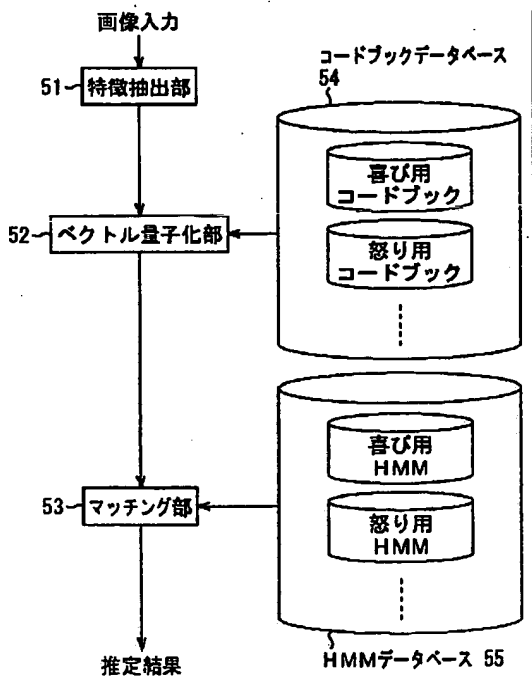
文生成部 4

【図６】



ユーザ感情情報更新部 8

【図７】



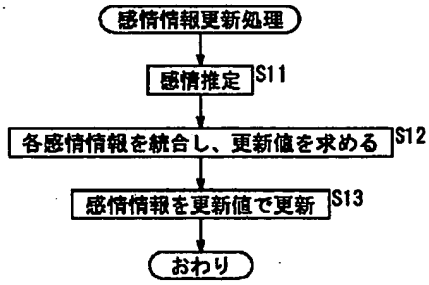
画像情報処理部 43

【図８】

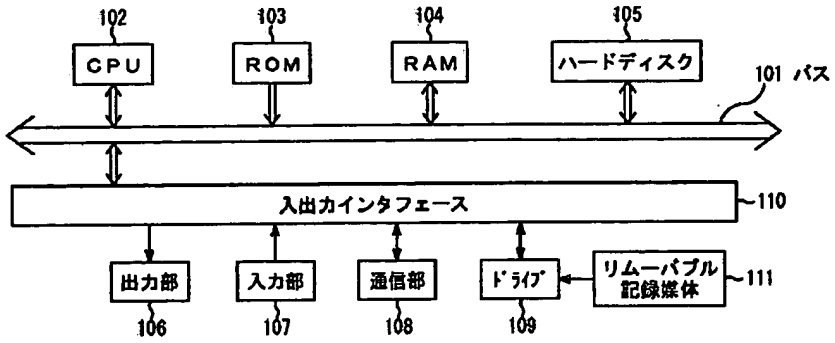
識別子	種類	値
P1	喜び	0.7
P2	怒り	0.2
P3	驚き	0.3
P4	悲しみ	0.1

感情情報

【図９】



【図１０】



コンピュータ

フロントページの続き

(72)発明者　田中　幸
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　横野　順
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　大江　敏生
東京都品川区北品川６丁目７番35号　ソニー株式会社内

Ｆターム(参考)　5D015 AA06 LL07 LL10
5D045 AB01 AB07 AB30
9A001 DZ11 FF03 HH17 HH18 HH33

Japanese Laid-open Patent
Laid-open Number: 2001-215993
Laid-open Date: August 10, 2001
Application Number: 2000-22225
Filing Date: January 31, 2000
Applicant: Sony Corporation

(54) [Title of the Invention] Interactive Processing device and method, and recording medium

(57) [Summary]

[Object] To make an interaction in a wide variety of forms depending on feelings of a user.

[Solving Means] A voice recognizing portion 2 recognizes voice of a user and extracts rhythm information of the voice. An interaction controlling portion 3 extracts concept information of words and phrases contained in the voice recognition results from the voice recognizing portion 2. An image inputting portion 6 captures an image of a face of the user and outputs face image information. A physiological information inputting portion 7 detects physiological information such as a pulse count of the user. Then, a user's feeling information updating portion 8 estimates a feeling of the user based on the rhythm information, concept information, face image information, and physiological information. The interaction controlling portion 3 and a sentence generating portion 4 generate an output sentence to be outputted to the user based

on the estimation results for the feeling.

[Scope of Claims]

[Claim 1] An interactive processing device for making an interaction with a user, characterized by comprising:

concept extracting means for extracting a concept of words and phrases inputted from a user;

feeling estimating means for estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and

output sentence generating means for generating an output sentence to be outputted to the user based on the feeling information.

[Claim 2] The interactive processing device according to claim 1, characterized in that the feeling estimating means estimates the feeling of the user based on the concept and the output sentence.

[Claim 3] The interactive processing device according to claim 1, characterized in that the feeling estimating means estimates the feeling of the user based on the concept and an image obtained by capturing an image of the user.

[Claim 4] The interactive processing device according to claim 1, characterized in that the feeling estimating means estimates the feeling of the user based on the concept and a physiological phenomenon of the user.

[Claim 5] The interactive processing device according to claim 1, further comprising sound processing means for processing a sound signal obtained from an outside,

characterized in that the feeling estimating means estimates the feeling of the user based on the concept and processing results of the sound processing means.

[Claim 6] The interactive processing device according to claim 1, further comprising voice recognizing means for recognizing voice of the user,

characterized in that the concept extracting means extracts a concept of words and phrases contained in voice recognition results for the voice of the user.

[Claim 7] The interactive processing device according to claim 6, characterized in that the feeling estimating means estimates the feeling of the user based on the concept and rhythm information of the voice of the user.

[Claim 8] The interactive processing device according to claim 1, characterized in that the output sentence generating means changes expression of the output sentence based on the feeling information.

[Claim 9] The interactive processing device according to claim 1, characterized in that the output sentence generating means changes the number of the output sentences based on the feeling information.

[Claim 10] The interactive processing device according to claim 9, characterized in that the output sentence means back-channel feedback.

[Claim 11] The interactive processing device according to claim 1, further comprising storage means for storing the feeling

information,

characterized in that the output sentence generating means generates the output sentence based on the feeling information stored in the storage means.

[Claim 12] The interactive processing device according to claim 1, characterized by further comprising output sentence outputting means for outputting the output sentence.

[Claim 13] The interactive processing device according to claim 12, characterized in that the output sentence outputting means outputs the output sentence as synthetic tone.

[Claim 14] The interactive processing device according to claim 13, characterized in that the output sentence outputting means controls a rhythm of the synthetic tone based on the feeling information.

[Claim 15] An interactive processing method for making an interaction with a user, characterized by comprising:

a concept extracting step of extracting a concept of words and phrases inputted from a user;

a feeling estimating step of estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and

an output sentence generating step of generating an output sentence to be outputted to the user based on the feeling information.

[Claim 16] A recording medium characterized by storing a program

for causing a computer to execute an interactive processing for making an interaction with a user, the program comprising:

a concept extracting step of extracting a concept of words and phrases inputted from a user;

a feeling estimating step of estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and

an output sentence generating step of generating an output sentence to be outputted to the user based on the feeling information.

[Detailed Description of the Invention]

[0001]

[Technical Field to which the Invention belongs] The present invention relates to an interactive processing device and method, and a recording medium, and more particularly to an interactive processing device and method, and a recording medium for allowing an interaction that reflects, for example, a feeling of a user.

[0002]

[Prior Art] In a so-called interactive system, when an input is made from a user, a response sentence corresponding to semantic contents of the input is generated to be outputted.

[0003]

[Problems to be solved by the Invention] Thus, in the conventional interactive system, irrespective of a feeling of a user, the same response sentence is outputted as long as the input has the same

semantic contents. As a result, the same interaction is made.

[0004] The present invention has been made in the light of such circumferences, and thus allows an interaction in a wide variety of forms depending on feelings of a user.

[0005]

[Means for solving the problem]

An interactive processing device according to the present invention is characterized by including: concept extracting means for extracting a concept of words and phrases inputted from a user; feeling estimating means for estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and output sentence generating means for generating an output sentence to be outputted to the user based on the feeling information.

[0006] The feeling estimating means may estimate the feeling of the user based on the concept and the output sentence.

[0007] Further, the feeling estimating means may estimate the feeling of the user based on the concept and an image obtained by capturing an image of the user.

[0008] Further, the feeling estimating means may estimate the feeling of the user based on the concept and a physiological phenomenon of the user.

[0009] The interactive processing device according to the present invention may further include sound processing means for processing

a sound signal obtained from an outside, in which the feeling estimating means may estimate the feeling of the user based on the concept and processing results of the sound processing means.

[0010] The interactive processing device according to the present invention may further include voice recognizing means for recognizing voice of the user, in which the concept extracting means may extract a concept of words and phrases contained in voice recognition results for the voice of the user.

[0011] The feeling estimating means may estimate the feeling of the user based on the concept and rhythm information of the voice of the user.

[0012] The output sentence generating means may change expression of the output sentence based on the feeling information.

[0013] The output sentence generating means may change the number of the output sentences based on the feeling information.

[0014] The output sentence means back-channel feedback.

[0015] The interactive processing device according to the present invention may further include storage means for storing the feeling information, in which the output sentence generating means generates the output sentence based on the feeling information stored in the storage means.

[0016] The interactive processing device according to the present invention may further include output sentence outputting means for outputting the output sentence.

[0017] The output sentence outputting means may output the output sentence as synthetic tone.

[0018] Further, the output sentence outputting means may control a rhythm of the synthetic tone based on the feeling information.

[0019] An interactive processing method according to the present invention is characterized by including: a concept extracting step of extracting a concept of words and phrases inputted from a user; a feeling estimating step of estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and an output sentence generating step of generating an output sentence to be outputted to the user based on the feeling information.

[0020] A recording medium according to the present invention is characterized by storing a program for causing a computer to execute an interactive processing for making an interaction with a user, the program comprising: a concept extracting step of extracting a concept of words and phrases inputted from a user; a feeling estimating step of estimating a feeling of the user based on the concept of the words and phrases inputted from the user to output feeling information expressing the feeling; and an output sentence generating step of generating an output sentence to be outputted to the user based on the feeling information.

[0021] In the interactive processing device and method, and the recording medium of the present invention, the concept of the words

and phrases inputted from the user is extracted, and the feeling of the user is estimated based on the extracted concept. Then, the output sentence to be outputted to the user is generated based on the feeling information obtained as a result of the estimation.

[0022]

[Embodiment Mode of the Invention] Fig. 1 shows an example of a configuration of an embodiment of an interactive system (the system means logical assembly of a plurality of apparatuses and it does not matter for the system whether or not apparatuses having respective configurations are accommodated in the same chassis) to which the present invention is applied.

[0023] A voice inputting portion 1, for example, is constituted by a microphone, an amplifier, and the like. The voice inputting portion 1 converts voice of a user into a sound signal as an electrical signal, amplifies the sound signal if necessary, and supplies the resultant sound signal to a voice recognizing portion 2.

[0024] The voice recognizing portion 2 acoustically processes the sound signal from the voice inputting portion 1, and recognizes the voice of the user based on the acoustic processing results. The voice recognition results are supplied to an interaction controlling portion 3. In addition, the voice recognizing portion 2 supplies rhythm information of the voice of the user, which is obtained by acoustically processing the sound signal to a user's feeling information updating portion 8.

[0025] The interaction controlling portion 3 generates contents of an output sentence to be outputted to the user as a response or the like to the voice recognition results from the voice recognizing portion 2 in consideration of the feeling information which expresses a feeling of the user and which is held (stored) in a user feeling information recording portion 9, and supplies content information expressing the contents to a sentence generating portion 4. In addition, the interaction controlling portion 3 extracts a concept of words and phrases contained in the voice recognition results from the voice recognizing portion 2, and words and phrases contained in the output sentence corresponding to content information generated by the interaction controlling portion 3 itself, and supplies concept information expressing the extracted concept to the user's feeling information updating portion 8.

[0026] The sentence generating portion 4, while taking the feeling information held by the user feeling information recording portion 9 into consideration, generates an output sentence in a text form corresponding to the content information from the interaction controlling portion 3, for example, and moreover generates a sound signal of a synthesis tone corresponding to the output sentence to supply the sound signal to a voice output portion 5.

[0027] The voice outputting portion 5, for example, is constituted by an amplifier, a speaker, and the like. The voice outputting portion 5 amplifies the sound signal from the sentence generating

portion 4 if necessary, and outputs the resultant sound signal through the speaker.

[0028] An image inputting portion 6, for example, is constituted by a lens, a CCD (Charge Coupled Device), an A/D converter, and the like. The image inputting portion 6 captures an image of a face or the like of the user and supplies face image information as digital data (image data) of a face image which is obtained as a result of the image capture to the user's feeling information updating portion 8.

[0029] A physiological information inputting portion 7, for example, is constructed by a pulse rate meter, a sensor for measuring an amount of perspiration, and a body temperature, and the like. The physiological information inputting portion 7 senses physiological phenomena such as a pulse rate, an amount of perspiration, and a body temperature of the user, and supplies the resultant physiological information to the user's feeling information updating portion 8.

[0030] The user's feeling information updating portion 8 estimates a feeling of the user based on the rhythm information of the voice of the user from the voice recognizing portion 2, the concept information of the words and phrases contained in the voice recognition results or the like from the interaction controlling portion 3, the face image information from the image inputting portion 6, and the physiological information from the physiological

information inputting portion 7. Moreover, the user's feeling information updating portion 8 updates feeling information held in the user feeling information recording portion 9 with feeling information which is obtained as a result of the estimation.

[0031] The user feeling information recording portion 9 holds feeling information in which feelings of joy, anger, surprise, and sorrow, for example, are expressed as feelings of the user in the form of numeric values falling within a predetermined range.

[0032] Next, a flow of basic processings in the interactive system of Fig. 1 will be described by referring to a flow chart of Fig. 2.

[0033] When an utterance is made by a user, the voice inputting portion 1 subjects a voice of the utterance to a voice inputting processing in Step S1, and outputs the resultant sound signal to the voice recognizing portion 2. That is, the voice inputting portion 1 converts the voice of the user into a sound signal as an electrical signal, amplifies the sound signal if necessary, and supplies the resultant sound signal to the voice recognizing portion 2.

[0034] In Step S2, the voice recognizing portion 2 recognizes the voice of the user based on the sound signal from the voice inputting portion 2, and supplies the voice recognition results to the interaction controlling portion 3. Moreover, the voice recognizing portion 2 extracts rhythm information of the voice of the user from the sound signal from the voice inputting portion 2, and supplies

the extracted rhythm information to the user's feeling information updating portion 8.

[0036] Thereafter, the process proceeds to Step S3 in which a preparation process for updating the feeling information held in the user feeling information recording portion 9 is executed.

[0036] That is, in Step S3, the interaction controlling portion 3 executes a feeling information updating interaction controlling processing for obtaining the above-mentioned concept information used to update the feeling information based on the voice recognition results or the like for the voice of the user from the voice recognizing portion 2, and supplies the resultant concept information to the user's feeling information updating portion 8. Moreover, in Step S3, the image inputting portion 6 executes an image inputting processing for capturing an image of a face of the user to obtain face image information, and supplies the resultant face image information to the user's feeling information updating portion 8. In addition, in Step S3, the physiological information inputting portion 7 executes a physiological information inputting processing for obtaining physiological information of the user, and supplies the resultant physiological information to the user's feeling information updating portion 8.

[0037] In Step S4, the user's feeling information updating portion 8 estimates a feeling of the user based on the rhythm information of the voice of the user from the voice recognizing portion 2, the

concept information from the interaction controlling portion 3, the face image information from the image inputting portion 6, and the physiological information from the physiological information inputting portion 7. Moreover, in Step S4, the user's feeling information updating portion 8 updates the feeling information held in the user feeling information recording portion 9 with the feeling information obtained as a result of the estimation.

[0038] Thereafter, in Step S5, the interaction controlling portion 3 executes a sentence generating interaction controlling processing for generating content information expressing contents of an output sentence to be outputted to the user as a response or the like to the voice recognition results from the voice recognizing portion 2 in consideration of the feeling information which expresses the feeling of the user and which is held (stored) in the user feeling information recording portion 9, and supplies the content information to the sentence generating portion 4.

[0039] Then, in Step S6, the sentence generating portion 4, while taking the feeling information held in the user feeling information recording portion 9 into consideration, generates an output sentence in a text form corresponding to the content information from the interaction controlling portion 3 (executes a sentence generating processing), and moreover generates a sound signal of the synthetic tone corresponding to the output sentence to supply the resultant sound signal to the voice outputting portion 5.

[0040] In Step S7, the voice outputting portion 5 executes a voice outputting processing for amplifying the sound signal from the sentence generating portion 4 to output the resultant sound signal through the speaker. Then, the operation is completed.

[0041] Note that in the above-mentioned case, in the interactive system, an output of the synthetic tone (hereinafter also referred to as an utterance in the interactive system as appropriate) is triggered by any utterance made by a user. Thus, the synthetic tone becomes a response to the utterance made by the user. However, in the interactive system, the utterance of the interactive system may also be triggered by any action other than an utterance made by the user.

[0042] That is, in the interactive system, for example, it is possible to make the utterance every predetermined period of time. In addition, for example, when the face image of the user is obtained in the image inputting portion 6 (including a case where the face image having a predetermined facial expression is obtained in addition to a case where the face image is simply obtained), or when predetermined physiological information is obtained in the physiological information inputting portion 7, it is possible to make the utterance. Moreover, for example, when a value of the feeling information held in the user feeling information recording portion 9 reaches a predetermined value or larger or smaller, it is also possible to make the utterance. In those cases, an

interaction is made such that the interactive system gives the utterance to the user, and the user responds to the utterance.

[0043] Next, Fig. 3 shows an example of a configuration of the voice recognizing portion 2 of Fig. 1.

[0044] The sound signal from the voice inputting portion 1 is supplied to an A/D (Analog Digital) converting portion 11. The A/D converting portion 11 converts the sound signal from the analog signal to a digital signal, and supplies the resultant sound data to a feature extracting portion 12. The feature extracting portion 12 subjects the sound data from the A/D converting portion 11 to the acoustic processing every arbitrary frame to extract a feature parameter such as a spectrum, a linear prediction coefficient, a cepstrum coefficient, a line spectrum pair, or MFCC (Mel Frequency Cepstrum Coefficient), and supplies the extracted feature parameter to a matching portion 13.

[0045] In addition, the feature extracting portion 12 supplies rhythm information such as an utterance speed, a pitch frequency, or a power which is obtained by subjecting the sound data to the acoustic processing to the user's feeling information updating portion 8. Note that a mora number or the like per frame, for example, can be used as the utterance speed.

[0046] The matching portion 13 recognizes the voice (input voice) of the user based on the feature parameter supplied from the feature extracting portion 12 by referring to an acoustic model database

14, a dictionary database 15, and a grammar database 16 if necessary.

[0047] That is, the acoustic model database 14 stores acoustic models expressing acoustic features such as individual phonemes and syllables in the language of the voice to be voice-recognized. Here, for example, HMM (Hidden Markov Model) and the like can be used as the acoustic models. The dictionary database 15 stores a word dictionary in which information related to pronunciations of words to be recognized. The grammar database 16 stores grammar rules that define the linkage (chain) between words registered in the word dictionary of the dictionary database 15. Here, for example, rules based on a contex-free grammar (CFG), HPSG (Head-driven Phrase Structure Grammar), or statistical word sequence probability (N-gram) can be used as the grammar rules.

[0048] The matching portion 13 connects the acoustic models stored in the acoustic model database 14 to construct the acoustic models (word models) of words by referring to the word dictionary of the dictionary database 15. Moreover, the matching portion 13 connects several word models by referring to the grammar rules stored in the grammar database 16. Then, the matching portion 13 recognizes the voice of the user based on the feature parameters through the HMM method or the like for example, using the word models which are connected in such a manner.

[0049] Then, the rhythm information as the voice recognition results given by the matching portion 13 is outputted in the form of a text

or a word graph for example, to the interaction controlling portion 3.

[0050] Fig. 4 shows an example of a configuration of the interaction controlling portion 3 of Fig. 1.

[0051] The voice recognition results for the voice of the user outputted from the voice recognizing portion 2 are supplied to a language processing portion 21. The language processing portion 21 processes the voice recognition results while referring to a thesaurus database 23, a language processing database 24, and a history database 25 if necessary, and supplies information on the meaning and concept expressed by the voice recognition results to the interaction processing portion 22.

[0052] That is, the thesaurus in which the words are classified in the form of hierarchical structure in accordance with their concepts is stored in the thesaurus database 23. The language processing portion 21 recognizes the concepts of the words contained in the voice recognition results by referring to the thesaurus.

[0053] Here, for example, "Word List by Semantic Principles" or the like which is published by The National Language Research Institute can be used as the thesaurus.

[0054] A word dictionary in which notation, necessary word class information, and the like of words are described, and syntax/meaning rules in which restrictions on the word sequence are described based on information of the words described on the word dictionary are

stored in the language processing database 24. The language processing portion 21 carries out the morphological analysis of the voice recognition results inputted thereto based on the word dictionary and the syntax/meaning rules. Moreover, the language processing portion 21 carries out the syntax analysis of the voice recognition results, and analysis of the semantic contents based on the morphological analysis results. Then, the language processing portion 21 outputs the results of the analysis of the concepts of the words constituting the voice recognition results thus obtained, and the semantic contents of the voice recognition results (hereinafter collectively referred to as the language processing results as appropriate) to the interaction processing portion 22.

[0055] Here, the language processing portion 21 can carry out the syntax analysis and semantic content analysis using the regular grammar, the contex-free grammar, the HPSG or the statistical word sequence probability.

[0056] In addition, the language processing portion 21 executes the processing with reference to the history database 25 as well if necessary. That is, the history (interaction history) of the interaction between the user and the interactive system is stored in the form of a set of voice recognition results of the voice uttered by the user and the interactive system's response to the utterance, or a set of an output of the interactive system and voice recognition

results of the voice uttered by the user for the output is stored in the history database 25. The language processing portion 21 allows the omission of the subject or the like in the voice recognition results, and analyzes the anaphoric expression or the like by referring to the interaction history. Thus, the language processing portion 21, for example, recognizes what a pronoun contained in the voice recognition results for the voice of the user concretely means.

[0057] Note that since the information stored in the thesaurus database 23 and the language processing database 24 is not basically updated, it can be said as so-called static information. On the other hand, since the interaction history stored in the history database 25 is updated by an interaction processing portion 22 which will be described later whenever the utterance is made by the user or the interactive system carries out any output to the user, it can be said as so-called dynamic information.

[0058] As described above, the language processing portion 21 extracts the concepts of the words (vocabularies) constituting the voice recognition results by referring to the thesaurus database 23. When the concept expresses a feeling, the language processing portion 21 supplies the information on the concept expressing the feeling as the concept information to the user's feeling information updating portion 8. That is, when the word belonging to the concept expressing a feeling such as "joy", "anger", "surprise", "sorrow",

"pain", "shamefulness" or "pleasure" on the thesaurus is contained in the voice recognition results, the language processing portion 21 supplies the concept information expressing that concept to the user's feeling information updating portion 8.

[0059] Note that the language processing portion 21 extracts the concept information of the words contained in the output of the interactive system stored as the interaction history in addition to the concept information of the words contained in the voice recognition results if necessary, and supplies the extracted information to the user's feeling information updating portion 8.

[0060] That is, the user's feeling information updating portion 8, as described above, estimates the feeling of the user. Then, the concept information of the words contained in the output of the interactive system as well as the concept information of the words contained in the voice recognition results are useful for the estimation in some cases. More specifically, for example, when such an utterance as to fool the user is made in the interactive system, it is anticipated that the user gets angry with the utterance. For this reason, the language processing portion 21 extracts the concept information as well of the words contained in the output of the interactive system stored as the interaction history by referring to the thesaurus, and supplies the extracted concept information together with the concept information of the words contained in the voice recognition results to the user's feeling

information updating portion 8.

[0061] The interaction processing portion 22, while referring to the history database 25 and a scenario database 26, generates contents of an output sentence to be outputted to the user as a response or the like to the voice recognition results for the voice of the user based on the language processing results from the language processing portion 21, and the feeling information which expresses a feeling of the user and which is held in the user feeling information recording portion 9, and supplies the content information expressing the contents to the sentence generating portion 4.

[0062] That is, the scenario database 26, for example, stores scenarios as rules of a pattern of an interaction with the user every task (topic). The interaction processing portion 22 basically determines the scenario to be used in an interaction with the user from the scenarios stored in the scenario database 26 based on the language processing results from the language processing portion 21, and generates the content information in accordance with the determined scenario.

[0063] More specifically, for example, the following scenario is stored for an objective-oriented task such as programming a VCR.

[0064] (action (Question (date, start_time, end_time, channel1)))

(date ???) #date

(start_time ???) #start time

(end_time ???) #end time

(channel ???) #channel

... (1)

[0065] Here, when the language processing results from the language processing portion 21 expresses a request for picture recording, the interaction processing portion 22 generates such content information as to instruct a user to set a date when picture recording is carried out, a time when the picture recording is started, a time when the picture recording is ended, and a channel for which the picture recording in this order in accordance with the scenario (1).

[0066] In addition, the following scenario is stored as the scenario for a free interaction (so-called idle talk) for example.

[0067] If X exist then speak (Y) #X: keyword, Y: response sentence

(Money What do you want) #(X Y)

(I want to eat something Do you feel hungry)

... (2)

[0068] Here, in accordance with the scenario (2), when a keyword of "Money" is contained in the language processing results from the language processing portion 21, the interaction processing portion 22 generates the content information as a question of "What do you want". In addition, when the keyword of "I want to eat something" is contained in the language processing results from the language processing portion 21, the interaction processing portion 22 generates the content information as a question of "Do

you feel hungry".

[0069] In addition, the interaction processing portion 22, for example, determines the scenario to be used based on not only the language processing results from the language processing portion 21, but also the feeling information held in the user feeling information recording portion 9. That is, for example, in a case where the language processing results from the language processing portion 21 express that the user said hello, when the feeling information expresses that "pleasure" or "happiness" is at a normal level, or when the feeling information expresses that "anger" or "irritation" is at a high level, the interaction processing portion 22 determines the use of the scenario for simply giving a reply of "hello" to the user. In addition, for example, in a case where the language processing results from the language processing portion 21 express that the user said hello, when the feeling information expresses that "pleasure" or "happiness" is at a high level, the interaction processing portion 22 determines the use of the scenario for putting a question of "did some happiness happen to you?" to the user.

[0070] Note that in addition to the scenarios, the general knowledge required to interact with the user is also stored in the scenario database 26. That is, for example, when the language processing results from the language processing portion 21 express that the user said hello, the information for instructing the interactive

system to give a reply to the user is stored as the general knowledge in the scenario database 26. In addition, for example, topics or the like used in idle talk are also stored as the general knowledge in the scenario database 26.

[0071] Moreover, the interaction processing portion 22 stores the language processing results from the language processing portion 21, the content information generated by the interaction processing portion 22 itself, the information related to the scenarios used to generate the content information, and the like as the interaction history in the history database 25.

[0072] In addition, the interaction processing portion 22 refers to the interaction history if necessary. Thus, for example, the interaction processing portion 22 also copes with a case where erroneous analysis of the voice recognition results, or erroneous semantic analysis thereof is detected later or the like.

[0073] Next, Fig. 5 shows an example of a configuration of the sentence generating portion 4 of Fig. 1.

[0074] The content information from the interaction controlling portion 3 is supplied to a text sentence generating portion 31. The text sentence generating portion 31 generates an output sentence in the text form corresponding (adapted) to the content information while referring to a dictionary database 34 and a generative grammar database 35 if necessary.

[0075] That is, a word dictionary in which word class information

of words, and information such as pronunciation and accents of the words are described is stored in the dictionary database 34. Templates of examples for an output sentence, declension rules for words necessary for generating an output sentence, and generating grammar rules such as restriction information of word order are stored in the generating grammar database 35. Then, the text sentence generating portion 31 selects the template corresponding to the content information, and selects necessary words from the word dictionary. Moreover, the text sentence generating portion 31, while suitably changing the ending or the like of a word, fits the words to the template by referring to the generating grammar rules, thereby generating an output sentence corresponding to the content information.

[0076] In addition, the feeling information held in the user feeling information recording portion 9 is also supplied to the text sentence generating portion 31. The text sentence generating portion 31 changes the expression of the output sentence based on the feeling information supplied thereto. That is, the templates which are identical in contents to one another but are different in expression from one another are stored in the generating grammar database 35. The text sentence generating portion 31 selects a template having a predetermined expression from such templates having the same contents, based on the feeling information. In addition, the text sentence generating portion 31 selects the words each having a

predetermined expression also for the words to be fitted to the template, based on the feeling information. Moreover, the text sentence generating portion 31 also changes the ending or the like of a word, based on the feeling information.

[0077] Thus, for example, when the feeling information expresses that "anger" or "sorrow" is at a high level, the text generating portion 31 generates an output sentence having a relatively polite expression. In addition, for example, when the feeling information expresses that "pleasure" or "joy" is at a high level, the text generating portion 31 generates an output sentence having a so-called rough expression.

[0078] Note that in addition to the method using templates, for example, a method based on a case structure or the like may also be adopted as a method of generating an output sentence.

[0079] After outputting an output sentence, the text sentence generating portion 31 carries out the morphological analysis, the syntax analysis, and the like to extract information necessary for voice rule synthesis carried out in a rule synthesizing portion 32 in a subsequent stage. Here, as for the information necessary for the voice rule synthesis, for example, there are the information for controlling a timing of the pause, an accent, and intonation, other rhythm information, rhythm information containing pronunciations of words, and the like.

[0080] The information obtained in the text sentence generating

portion 31 is supplied to the rule synthesizing portion 32. The rule synthesizing portion 32 generates the sound data (digital data) of synthetic tone corresponding to the output sentence generated in the text sentence generating portion 31 using a phoneme database 36.

[0081] That is, for example, phoneme data is stored in the form of CV (Consonant, Vowel), VCV, CVC or the like in the phoneme database 36. The rule synthesizing portion 32 combines the necessary phoneme data based on the information from the text sentence generating portion 31. The rule synthesizing portion 32 suitably applies the pause, the accent, the intonation, and the like to generate sound data of the synthetic sound corresponding to the output sentence generated in the text sentence generating portion 31.

[0082] In addition, the feeling information held in the user feeling information recording portion 9 is supplied to the rule synthesizing portion 32. The rule synthesizing portion 32 controls the rhythm information such as the pause, the accent, the intonation, the utterance speed, and a pitch frequency applied to the combined phoneme data based on the feeling information. That is, for example, when the feeling information expresses that the user is excited, the rule synthesizing portion 32 generates the sound data of a synthetic sound having a slow and calm tune. In addition, for example, when the feeling information expresses that the user seems pleasant, the rule synthesizing portion 32 also generates the sound data of

a synthetic sound having a pleasant tune.

[0083] Note that the details of the relationship between the feeling and the voice are described in Maekawa: "Transmission of Paralanguage Information Through Voice; From a Viewpoint of Linguistics", Proceedings of the (1997) Autumn Meeting of the Acoustical Society of Japan, 1-3-10, pp. 381 to 384, September, 1997, or the like.

[0084] The sound data of the synthetic sound obtained in the rule synthesizing portion 32 is supplied to a D/A (Digital Analog) converting portion 33, and the D/A converting portion 33 converts the sound data supplied thereto into the sound signal as an analog signal. The resultant sound signal is supplied to the voice outputting portion 5 from which a synthetic sound corresponding to the output sentence generated in the text sentence generating portion 31 is in turn outputted.

[0085] Next, Fig. 6 shows an example of a configuration of the user's feeling information updating portion 8 of Fig. 1.

[0086] The rhythm information outputted by the voice recognizing portion 2, the concept information outputted by the interaction controlling portion 3, the face image information outputted by the image inputting portion 6, and the physiological information outputted by the physiological information inputting portion 7 are supplied to a rhythm information processing portion 41, a concept information processing portion 42, an image information processing portion 43, and a physiological information processing portion 44,

respectively.

[0087] The rhythm information processing portion 41 processes the rhythm information supplied thereto to estimate a feeling of the user, and outputs the feeling information as the estimation results to an update processing portion 45.

[0088] Note that for example, a method or the like described in JP 10-55194 A can be used as a method of estimating a feeling of a user based on the rhythm information of a voice of the user.

[0089] The concept information processing portion 42 processes the concept information supplied thereto to estimate a feeling of the user, and outputs feeling information as the estimation results to the update processing portion 45. That is, the concept information processing portion 42 measures an appearance frequency at which words belonging to the concept expressing feelings such as "joy" and "anger" appears in the interaction between the user and the interactive system, based on the concept information. Then, the concept information processing portion 42 estimates a feeling of the user based on the appearance frequency, and outputs feeling information as the estimation results to the update processing portion 45.

[0090] The image information processing portion 43 processes the face image information supplied thereto to estimate a feeling of the user, and outputs feeling information as the estimation results to the update processing portion 45.

[0091] That is, Fig. 7 shows an example of a configuration of the image information processing portion 43 of Fig. 6.

[0092] The face image information is supplied to a feature extracting portion 51, and the feature extracting portion 51 extracts feature quantities of the face image information. That is, the feature extracting portion 51, for example, wavelet-converts the face image information to obtain a feature vector having a coefficient expressing a spatial frequency component as its component, and supplies information on the resultant feature vector to a vector quantization portion 52.

[0093] The vector quantization portion 52 vector-quantizes the feature vector from the feature extracting portion 51 in accordance with a code book stored in a code book database 54 to obtain a one-dimensional symbol (column).

[0094] That is, the code books which are obtained by carrying out learning using images of a face in such feeling as joy, anger, surprise, and sorrow are stored in the code book database 54. Note that in this example, in order to enhance the quantization precision, the code books for individual feelings such as a joy code book, and an anger code book are created to be stored in the code book database 54.

[0095] Then, the vector quantization portion 52 vector-quantizes the feature vector from the feature extracting portion 51 in accordance with the code books for individual feelings stored in

the code book database 54 to obtain a symbol (a code assigned to a code vector of the code book), and outputs the resultant symbol to a matching portion 53. Consequently, the symbol as the vector quantization results for individual feelings is supplied to the matching portion 53.

[0096] The matching portion 53 carries out the matching in order to determine which of faces in a joyful feeling, an angry feeling, a surprising feeling, and a sorrowful feeling, for example, corresponds to the face image information using the symbol from the vector quantization portion 52 by referring to an HMM database 55.

[0097] That is, models (HMM) about the faces in the individual feelings which are obtained by carrying out the learning using the images of the face in such feeling as joy, anger, surprise, and sorrow are stored in the HMM database 55.

[0098] Then, the matching portion 53 obtains the model with the highest probability at which the symbol series obtained from the vector quantization portion 52 is observed by utilizing a Viterbi method. Moreover, the matching portion 53 estimates the feeling corresponding to that model as a feeling of the user, and outputs the feeling information as the estimation results to the update processing portion 45.

[0099] Here, the calculation for the probability at which the symbol series obtained from the vector quantization portion 52 is observed

is carried out every feeling in the matching portion 53. That is, for example, the calculation for the probability of the observation of the symbol series obtained by carrying out the vector quantization using the joy code book, for example is carried out using the HMM (joy HMM) obtained by carrying out the leaning using the image of the face in the joyful feeling. In addition, the calculation for the probability of the observation of the symbol series obtained by carrying out the vector quantization using the anger code book, for example is carried out using the HMM (anger HMM) obtained by carrying out the leaning using the image of the face in the angry feeling.

[0100] Note that the details of the method of estimating a feeling from face image information in the manner as described above are described in, for example, Sakaguchi, Ohya, and Kishino: "Facial Expression Recognition from Image Sequence using Hidden Markov Model", The Journal of the Institute of Television Engineers of Japan, Vol. 49, No. 8, pp. 1060 to 1067, August, 1995, or the like.

[0101] A method described in Sakaguchi, Morishima: "Real-time Basic Facial Expression Recognition based on Spatial Frequency Information", Proceedings of the 2-nd Symposium on Intelligence Information Media, pp. 75-82, December, 1996, or the like can also be adopted as the method of estimating a feeling from face image information.

[0102] Referring back to Fig. 6, the physiological information

processing portion 44 processes the supplied physiological information to estimate a feeling of the user, and outputs the feeling information as the estimation results to the update processing portion 45. Here, as for a method of estimating a feeling of the user from the physiological information, for example, there is known a method in which a function expressing a correlation between feelings and physiological information such as a pulse count or an amount of perspiration is statistically obtained in advance, and a feeling of a user is estimated using the resultant function, or the like.

[0103] The update processing portion 45 obtains a final update value with which the feeling information held in the user feeling information recording portion 9 is to be updated using synthetically the feeling information from the rhythm information processing portion 41, the concept information processing portion 42, the image information processing portion 43, and the physiological information processing portion 44, and updates the feeling information held in the user feeling information recording portion 9 with the resultant update value. That is, the update processing portion 45 executes, for example, weighted-addition and normalization of the feeling information corresponding to the respective feelings from the rhythm information processing portion 41, the concept information processing portion 42, the image information processing portion 43, and the physiological information processing portion 44 to derive final feeling information corresponding to the respective feelings.

Then, the update processing portion 45 updates the feeling information held in the user feeling information recording portion 9 with the final feeling information.

[0104] Here, Fig. 8 shows the feeling information held in the user feeling information recording portion 9. For the feeling information corresponding to the respective feelings, the strength of the feeling is expressed by an actual number from 0 to 1, for example. Thus, the larger number means the stronger feeling (the smaller the actual number, the weaker the feeling is). The update processing portion 45 updates the actual number as such feeling information every feeling.

[0105] Next, processings (feeling information updating processings) in the user's feeling information updating portion 8 of Fig. 6 will be described by referring to a flow chart shown in Fig. 9.

[0106] First of all, in Step S11, the rhythm information processing portion 41, the concept information processing portion 42, the image information processing portion 43, and the physiological information processing portion 44 estimate a feeling of the user in the manner as described above, and output the feeling information as the estimation results to the update processing portion 45.

[0107] In Step S12, the update processing portion 45 obtains the final update value with which the feeling information held in the user feeling information recording portion 9 is to be updated using synthetically the feeling information from the rhythm information

processing portion 41, the concept information processing portion 42, the image information processing portion 43, and the physiological information processing portion 44. Then, the process proceeds to Step S13. In Step S13, the update processing portion 45 updates the feeling information held in the user feeling information recording portion 9 with the update value. Thus, the processings are completed.

[0108] Next, the above-mentioned series of processings can be executed by use of dedicated hardware or software. When the series of processings are executed by use of the software, a program constituting the software is installed in a general purpose computer or the like.

[0109] Then, Fig. 10 shows an example of a configuration of an embodiment of the computer in which the program for executing the above-mentioned series of processings is installed.

[0110] The program can be previously recorded on a hard disc 105 or a ROM 103 as a recording medium incorporated in the computer.

[0111] Alternatively, the program can be temporarily or permanently stored (recorded) in a removal recording medium 111 such as a floppy disc, a CD-ROM (Compact Disc Read Only Memory), an MO (Magneto Optical) disc, a DVD (Digital Versatile Disc), a magnetic disc, or a semiconductor memory. Such a removal recording medium 111 can be presented as so-called package software.

[0112] Note that the program can be installed in the computer from

the removal recording medium 111 as described above, and in addition, can be wirelessly transferred from a download site to the computer through an artificial satellite for digital satellite broadcasting, or can be wiredly transferred to the computer through a network such as a LAN (Local Area Network) or the Internet. In the computer, the program which is transferred thereto in such a manner can be received by a communication portion 108 and can be installed in the hard disc 105 incorporated therein.

[0113] The computer incorporates a CPU (Central Processing Unit) 102. An I/O interface 110 is connected to the CPU 102 through a bus 101. When a command is inputted to the CPU 102 through manipulation of an input portion 107 constituted by a keyboard, a mouse or the like by the user, the CPU 102 executes the program stored in a ROM (Read Only Memory) 103 in accordance with the command. Besides, the CPU 102 loads a program stored in the hard disc 105, a program which is transferred through a satellite or network communication and then received by the communication portion 108 to be installed in the hard disc 105, or a program which is read out from the removal recording medium 111 inserted to a drive 109 to be installed in the hard disc 105 into a RAM (Random Access Memory) 104 to execute the program. Thus, the CPU 102 executes processings in accordance with the above-mentioned flow chart, or processings carried out with the configuration of the above-mentioned block diagram. Then, the CPU 102, if necessary, outputs the processing

results with an output portion 106 constituted by a liquid crystal display (LCD), a speaker or the like, or transmits the processing results from the communication portion 108, or records the processing results on the hard disc 105 through the I/O interface 110.

[0114] Here, in this specification, the processing steps described in the program with which the computer is instructed to execute various kinds of processings are not necessarily executed in a time series manner in accordance with the flow of the flow chart. Thus, the processing steps also include processings which are executed in parallel or individually (e.g., parallel processings or object-based processings).

[0115] In addition, the program may be executed by a single computer or may be distributedly executed by a plurality of computers. Moreover, the program may also be transferred to a remote computer to be executed by the remote computer.

[0116] As described above, since a feeling of the user is estimated based on at least the concept of words and phrases contained in the voice recognition results for the voice of the user, the feeling of the user can be estimated with relatively high precision. Moreover, since a feeling of the user is estimated based on the rhythm information, the face image information, and the physiological information in addition thereto, the feeling of the user can be estimated more precisely. Furthermore, since the output sentence is generated based on such feeling estimation results, the output

sentence can be presented to the user in a wide variety of forms depending on the feelings of the user.

[0117] In this embodiment, the voice recognition is carried out for the sound (voice) inputted to the voice inputting portion 1. However, the voice recognition may not be carried out for the sound inputted to the voice inputting portion 1. In this case, for example, the sound may be detected as a sound of tapping on a desk with user's fingers, or a sound of breathing of a user, and a feeling of the user may also be estimated based on the detection results. That is, for example, when it is continuously detected that a desk is tapped, it is possible to estimate that the user gets angry. In addition, for example, when it is detected that the user breathes hard, it is possible to estimate that the user is excited. In this case, it is possible to apply such an ad hoc update rule as to increase the value of the feeling information expressing "anger" or "excitement" based on such estimation results.

[0118] Moreover, in the interaction controlling portion 3, the number of times of utterance to the user can be changed by controlling the number of times of generation of the output sentence in correspondence to a user's feeling. More specifically, for example, when the user seems pleasant, for example, the number of times of the back-channel feedback is increased, and in addition thereto, the number of times of the utterance from the interactive system is increased. Thus, it is possible to positively make the interaction

with the user. In addition, for example, when the user seems sorrowful, the number of times of the utterance from the interactive system is decreased. Thus, it is possible to prevent the user from feeling troublesomeness.

[0119] In addition, in this embodiment, the voice from the user is recognized, and the utterance is made in response to the voice recognition results. Besides, for example, a response may also be made to a sentence inputted through the user's manipulation of a keyboard.

[0120] Moreover, in this embodiment, the response or the like to the user is outputted in the form of the synthetic tone. In addition thereto, for example, the response or the like to the user may also be displayed in the form of a text or the like.

[0121] In addition, the present invention can be used as, for example, a user interface between virtual characters displayed on a display device, or a physical robot or the like and the user. In this case, as the response or the like to the user, in addition to the output of the synthetic tone as described above, a display state of virtual characters is changed or a robot is made to carry out a predetermined operation. Thus, it is possible to realize a multi-modal interface.

[0122]

[Effects of the Invention] According to the interactive processing device and method, and the recording medium of the present invention, the concept of the words and phrases outputted from the user is

extracted, and a feeling of the user is estimated based on the extracted concept. Then, the output sentence to be outputted to the user is generated based on the resultant feeling information. Consequently, the interaction can be made in a wide variety of forms depending on feelings of the user, for example.

[Brief Description of the Drawings]

[Fig. 1] A block diagram showing an example of a configuration of an embodiment of an interactive system to which the present invention is applied.

[Fig. 2] A flow chart for explaining processings in the interactive system of Fig. 1.

[Fig. 3] A block diagram showing an example of a configuration of a voice recognizing portion 2 of Fig. 1.

[Fig. 4] A block diagram showing an example of a configuration of an interaction controlling portion 3 of Fig. 1.

[Fig. 5] A block diagram showing an example of a configuration of a sentence generating portion 4 of Fig. 1.

[Fig. 6] A block diagram showing an example of a configuration of a user's feeling information updating portion 8 of Fig. 1.

[Fig. 7] A block diagram showing an example of a configuration of an image information processing portion 43 of Fig. 6.

[Fig. 8] A diagram showing feeling information.

[Fig. 9] A flow chart for explaining processings in a user's feeling information updating portion 8 of Fig. 6.

[Fig. 10] A block diagram showing an example of a configuration of an embodiment of a computer to which the present invention is applied.

[Description of Reference Numerals]

1, voice inputting portion; 2, voice recognizing portion; 3, interaction controlling portion; 4, sentence generating portion; 5, voice outputting portion; 6, image inputting portion; 7, physiological information inputting portion; 8, user's feeling information updating portion; 9, user feeling information recording portion; 11, A/D converting portion; 12, feature extracting portion; 13, matching portion; 14, acoustic model database; 15, dictionary database; 16, grammar database; 21, language processing portion; 22, interaction processing portion; 23, thesaurus database; 24, language processing database; 25, history database; 26, scenario database; 31, text sentence generating portion; 32, rule synthesizing portion; 33, D/A converting portion; 34, dictionary database; 35, generating grammar database; 36, phoneme database; 41, rhythm information processing portion; 42, concept information processing portion; 43, image information processing portion; 44, physiological information processing portion; 51, feature extracting portion; 52, vector quantization portion; 53, matching portion; 54, code book database; 55, HMM database; 101, bus; 102, CPU; 103, ROM; 104, RAM; 105, hard disc; 106, output portion; 107, input portion; 108, communication portion;109, drive; 110, I/O

interface; and 111, removal recording medium

FIG. 1

INTERACTIVE SYSTEM

1 VOICE INPUTTING PORTION

2 VOICE RECOGNIZING PORTION

3 INTERACTION CONTROLLING PORTION

4 SENTENCE GENERATING PORTION

5 VOICE OUTPUTTING PORTION

6 IMAGE INPUTTING PORTION

7 PHYSIOLOGICAL INFORMATION INPUTTING PORTION

8 USER'S FEELING INFORMATION UPDATING PORTION

9 USER FEELING INFORMATION RECORDING PORTION

SYNTHETIC TONE

FIG. 2

START

S1 VOICE INPUT PROCESSING

S2 VOICE RECOGNITION PROCESSING

S3 PREPARATION FOR UPDATE OF FEELING INFORMATION

FEELING INFORMATION UPDATING INTERACTION CONTROL PROCESSING

IMAGE INPUT PROCESSING

PHYSIOLOGICAL INFORMATION INPUT PROCESSING

S4 UPDATE FEELING INFORMATION

S5 SENTENCE GENERATING INTERACTION CONTROL PROCESSING

S6 SENTENCE GENERATION PROCESSING

S7 VOICE OUTPUT PROCESSING

END

FIG. 3

VOICE INPUT

2 VOICE RECOGNIZING PORTION

11 A/D CONVERTING PORTION

12 FEATURE EXTRACTING PORTION

13 MATCHING PORTION

14 ACOUSTIC MODEL DATABASE

15 DICTIONARY DATABASE

16 GRAMMAR DATABASE

RHYTHM INFORMATION

VOICE RECOGNITION RESULTS

FIG. 4

VOICE RECOGNITION RESULTS

(TEXT/WORD GRAPH)

3 INTERACTION CONTROLLING PORTION

21 LANGUAGE PROCESSING PORTION

CONCEPT INFORMATION

(CONCEPT/SEMANTICS)

22 INTERACTION PROCESSING PORTION

FEELING INFORMATION

CONTENT INFORMATION

23 THESAURUS DATABASE

24 LANGUAGE PROCESSING DATABASE

SYNTAX/SEMANTIC RULE WORD DICTIONARY

25 HISTORY DATABASE

INTERACTION HISTORY

26 SCENARIO DATABASE

SCENARIO

KNOWLEDGE

FIG. 5

RESPONSE CONTENTS

4 SENTENCE GENERATING PORTION

31 TEXT SENTENCE GENERATING PORTION

32 RULE SYNTHESIZING PORTION

33 D/A CONVERTING PORTION

SOUND DATA

34 DICTIONARY DATABASE

35 GENERATING GRAMMAR DATABAE

36 PHONEME DATABASE

FIG. 6

RHYTHM INFORMATION

CONCEPT INFORMATION

FACE IMAGE INFORMATION

PHYSIOLOGICAL INFORMATION

8 USER'S FEELING INFORMATION UPDATING PORTION

41 RHYTHM INFORMATION PROCESSING PORTION

42 CONCEPT INFORMATION PROCESSING PORTION

43 IMAGE INFORMATION PROCESSING PORTION

44 PHYSIOLOGICAL INFORMATION PROCESSING PORTION

45 UPDATE PROCESSING PORTION

UPDATE VALUE

FIG. 7

IMAGE INPUT

43 IMAGE INFORMATION PROCESSING PORTION

51 FEATURE EXTRACTING PORTION

52 VECTOR QUANTIZATION PORTION

53 MATCHING PORTION

ESTIMATION RESULTS

54 CODE BOOK DATABASE

JOY CODE BOOK

ANGER CODE BOOK

55 HMM DATABASE

JOY HMM

ANGER HMM

FIG. 8

FEELING INFORMATION

IDENTIFIER

KIND

VALUE

JOY

ANGER

SURPRISE

SORROW

FIG. 9

FEELING INFORMATION UPDATING PROCESSING

S11 ESTIMATE FEELING

S12 COMBINE INDIVIDUAL FEELING INFORMATION AND DERIVE UPDATE VALUE

S13 UPDATE FEELING INFORMATION WITH UPDATE VALUE

END

FIG. 10

COMPUTER

101 BUS

105 HARD DISC

106 OUTPUT PORTION

107 INPUT PORTION

108 COMMUNICATION PORTION

109 DRIVE

110 I/O INTERFACE

111 REMOVAL RECORDING MEDIUM

PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2001-215993**
(43)Date of publication of application : **10.08.2001**

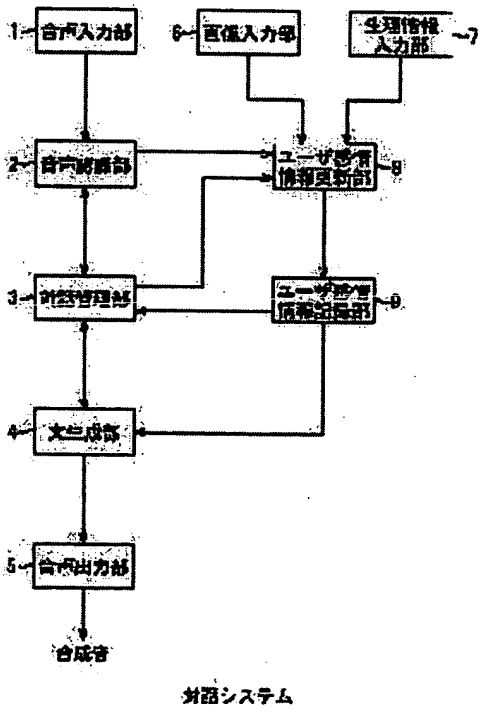
(51)Int.CI. G10L 15/22
G06F 3/16
G10L 13/00
G10L 15/10

(21)Application number : **2000-022225**
(22)Date of filing : **31.01.2000**

(71)Applicant : **SONY CORP**
(72)Inventor : **ASANO KOJI**
AOYANAGI SEIICHI
TANAKA MIYUKI
YOKONO JUN
OE TOSHIO

(54) DEVICE AND METHOD FOR INTERACTIVE PROCESSING AND RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To conduct interactive operations having rich variations depending on the feeling condition of a user.
SOLUTION: In a voice recognition section 2, user's voice is recognized and phoneme information of the voice is extracted. In an interactive control section 3, conceptual information of the words and the phrases included in the voice recognition result obtained by the section 2 is extracted. An image inputting section 6 photographs the face of the user and outputs face image information. In a physiological information inputting section 7, physiological information such as the pulse rate of the user is detected. Then, a user feeling information updating section 8 estimates the feeling of the user based on the phoneme, the conceptual, the face image and the physiological information. In the section 3 and a sentence generating section 4, an output sentence is generated and outputted to the user based on the estimated result of the feeling.



対話システム

LEGAL STATUS
[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]